Marginal likelihood in state-space models

Francke, M.K.

2006

document version
Publisher's PDF, also known as Version of record

MARGINAL LIKELIHOOD IN STATE-SPACE MODELS

THEORY AND APPLICATIONS

Cover design: Arjan Gras, Gras Communicatiebureau, 's Hertogenbosch.

Printing of this thesis was financially supported by OrtaX.

VRIJE UNIVERSITEIT

MARGINAL LIKELIHOOD IN STATE-SPACE MODELS
THEORY AND APPLICATIONS

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor aan
de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof.dr. T. Sminia,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de faculteit der Economische Wetenschappen en Bedrijfskunde
op dinsdag 7 maart 2006 om 13.45 uur
in de aula van de universiteit,
De Boelelaan 1105

door

Marcus Kornelis Francke

geboren te Middelburg

promotor: prof.dr. S.J. Koopman
copromotor: dr. A.F. de Vos

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Review

This thesis starts with chapters on econometric theory and ends with practice. The road through time was the other way around. The motivation was a real econometric problem: the development of econometric models to predict the market value of all houses in the city of Amsterdam. The main methodology used in this application is the Kalman filter. Some questions concerning the so called diffuse Kalman filter and the associated diffuse likelihood led to interest in marginal likelihood. The discovery that the concept of marginal likelihood is the key to solve the "unit root problem" shifted interest to this classic econometric problem. The similarity between solutions on the basis of Bayesian and classical marginal likelihood approaches even led into the deep waters of theory about noninformative priors. This section describes the journey, while the next section sets out the thesis.

Market values of houses are used by the local and central government and water boards for tax purposes. The market value is defined in the law WOZ (Waardering Onroerende Zaken, article 17, part 2) as "the value when full and unencumbered ownership is transferred and the buyer can take possession of that immediately and completely". There is high quality data available for developing models, several thousands transactions a year over a long time period, for which sales prices have become available together with the values of the property characteristics which might serve as explanatory variables.

The "Dienst Belastingen Gemeente Amsterdam" (Amsterdam Tax Authorities Office) started a research team in 1993 to develop valuation models. In the first years the model was used as a second opinion. Real estate appraisers did the actual work of valuing real estate. In a later period the models were used to handle objections against valuations made in the traditional way. Needham, Francke, and Bosma (1998) provide an overview of the practical and research results in these years. From the year 2001 the situation has turned the other way. The model was used for valuation and the appraiser validated the property value predicted by the model. Now the model is operational for mass appraisal in a number of mainly larger cities in the

Netherlands.

Due to the fact that a large part of the housing stock in Amsterdam is let at low and/or subsidised rents, the relative number of transactions that meet the conditions of the law WOZ is low, only 5,000 usable transactions on a housing stock of 380,000. In a number of neighborhoods dwellings are rarely ever sold. For a valuation in period t transactions from previous and following periods have to be used in order to provide a sound basis for valuation. The natural thing is to set-up a time series model. A hierarchical trend model was specified in which cluster-level, general trends, and specific characteristics play a role. Examples of clusters are districts and house types. The general trend, and the cluster-level as deviations from the general trend, are modeled as stochastic trends on a monthly basis. The deterministic part is a nonlinear function of the characteristics.

The model, formulated in state-space form, can be estimated by the Kalman filter. An efficient estimation procedure for dealing with the large number of observations is to decompose the original model into two parts, that are treated differently. The first part is ordinary least squares on data in deviation from means. This step provides a prior for coefficients to be used in the second step, that consists of the Kalman filter. In this step estimates of the trends and the parameters are obtained. The procedure exploits and illustrates the Bayesian interpretation of a Kalman filter.

The initial condition of the hierarchical trend model is unknown, due to the fact that the model contains regression parameters and nonstationary components. The diffuse Kalman filter provides an efficient way to cope with the situation of an unknown initial condition. The filter produces a profile or diffuse likelihood, depending on whether the initial condition is treated as a fixed or random variable. The likelihood is used for inference on parameters in the system matrices of the state-space model. In the literature on Kalman filtering almost no motivation is provided for what type of likelihood must be used. I became interested when I realized that the diffuse likelihood depends on the specific state-space representation of the same model, and that the difference in diffuse likelihood between two specifications may depend on the parameters of interest.

A justification for the use of the diffuse likelihood was found in another part of the statistical literature, concerning classical marginal likelihood in the general linear model $y = X\beta + u$, with $u \sim N(0, \sigma^2\Omega(\theta))$. Marginal likelihood concerns inference on the parameters θ in the covariance matrix. The regression and scale parameters are regarded as nuisance. The marginal likelihood is the likelihood of a transformation of the data y such that the transformed data do not depend on β and σ. The concept of marginal likelihood was introduced by Kalbfleisch and Sprott (1970). The use of the classical marginal likelihood is limited to location and scale parameters and some other applications, which may explain that it remained relatively unknown. However, in state-space models it can be used and the marginal likelihood can easily be calculated from the diffuse likelihood. Unlike the diffuse likelihood, the marginal likelihood is invariant to regular transformations of the explanatory variables.

The main difference with the usual profile likelihood is the term $|\sigma^{-2}X'\Omega^{-1}X|^{-1/2}$. Consequently inference on σ and θ differs. Estimation based on marginal likelihood is to be preferred since it adjusts for the evidence on θ in the part of the data that is a linear function of X, which is pseudo-information. The question is whether there is any difference in inference based on profile and marginal likelihood in practice. In many cases the difference between profile and marginal likelihood based inference is small, especially for large sample sizes. An example where the difference matters even asymptotically, is the famous econometric "unit root problem" in the linear model with first order autoregressive disturbances. Unlike the profile likelihood, the marginal likelihood is well behaved when the autoregressive parameter $\rho = 1$: finite, nonzero and continuous for $\rho \uparrow 1$. Unit root tests based on the marginal likelihood appear to be more powerful than other well-known and popular tests. Power functions almost coincide with the power envelope, even in small samples, although no uniformly most powerful test exists.

The term marginal likelihood is not only used in classical, but also in Bayesian statistics. In Bayesian statistics the nuisance parameters are integrated out. Unlike the prior following from Jeffreys' rule, the independence Jeffreys' prior appears to establish proportionality between both classical and Bayesian marginal likelihood in the general linear model. It is argued that there is a strong case to use this prior, or even classical marginal likelihood directly. Consequently classical tests and Bayes factors can be based on the same marginal likelihood ratio.

In the context of hypothesis testing there are major differences between classical and Bayesian inference. However, in case of a marginal likelihood depending on 1 parameter and a monotone marginal likelihood ratio, Bayesian posterior odds and classical marginal likelihood ratio tests, use the data exactly in the same way. The only difference is the size: in a classical study a predetermined value and in the Bayesian context following from prior considerations. In the Bayesian analysis it is also possible to compute a "p-value". This statistic provides all relevant information from the data, and facilitates the discussion with classical statisticians.

The proportionality of classical and Bayesian marginal likelihood, and the the same use of the data in hypothesis tests, is applied to the unit root example. A number of noninformative priors $\pi(\beta, \sigma^2|\rho)$ is proposed in literature, most of them leading to a posterior that is zero in the unit root. The independence Jeffreys' does not have this problem. The classical and Bayesian unit root tests are almost indistinguishable, because the marginal likelihood ratio is almost monotone.

This thesis concerns different worlds. First of all it concerns hedonic price models in real estate economics. The techniques used to estimate these models come from the Kalman filtering literature. The justification of the diffuse likelihood was found in the classical literature on marginal likelihood. Similarities between classical and Bayesian inference based on marginal likelihood are shown for the general linear model. Bayesian and classical statistics can be compared with different religions, almost without mutual communication. This thesis partly tries to connect the different worlds, and provides some material that can serve as a bridge for mutual understanding.

1.2 Overview of the thesis

The set-up of this thesis is as follows. In chapter 2 the concept of the classical marginal is introduced for the general linear model and is applied to the linear model with first order autoregressive disturbances. The null hypothesis of a unit root is tested by marginal likelihood ratio tests. The asymptotic distribution of the marginal likelihood ratio is derived and the power functions of the marginal likelihood ratio tests are compared to other unit root tests

In chapter 3 it is shown that Bayesian and classical marginal likelihood are proportional when the independence Jeffreys' prior is used. Further, in the case of a monotone marginal likelihood ratio depending on only one parameter, it is shown that the marginal likelihood ratio test and the Bayesian posterior odds test have the same power function. These results are applied to the "unit root problem". The relation between the classical marginal likelihood ratio tests and the posterior odds tests are studied for different priors of the autoregressive parameter ρ and for different sample sizes.

Chapter 4 deals with inference on parameters in the system matrices of state-space models with diffuse initial conditions. The different likelihood concepts – profile, diffuse, and (concentrated) marginal – are compared with each other in the context of estimation, testing and model comparison. A motivation for the use of the marginal (and so in many cases the diffuse) likelihood is provided.

Chapter 5 concerns the efficient estimation of the hierarchical trend model, in which cluster-level trends, general trends, and specific characteristics play a role. A new procedure to implement a state-space model for repeated measurements is provided.

Chapter 6 presents estimation results for the hierarchical trend model. This model is used for property valuation and determining local price indices. Price indices based on the hierarchical trend model are compared to a standard hedonic index and an index based on weighted median selling prices published by national brokerage organization. It is shown that, especially for small housing market segments, the hierarchical trend model produces price indices, that are more accurate, detailed, and up-to-date.

Chapter 2

Marginal likelihood and unit roots

Abstract[1]

We develop new tests for the hypothesis of unit roots that are based on the marginal likelihood of the general linear model. The marginal likelihood allows the incorporation of invariance arguments in the likelihood function. It turns out that marginal likelihood tests for unit roots appear to be more powerful than other unit root tests. For some basic models power functions almost coincide with the power envelopes, even in small samples. General correlation structures can be incorporated, either by standard likelihood procedures or by adjustments of the test statistics on the basis of asymptotic distributions.

2.1 Introduction

In the econometrics literature on time series a variety of procedures has appeared for testing the null hypothesis of a unit root. Different specifications as well as different methods of inference have been used. Müller and Elliott (2003) show that most differences come down to a different way of handling the initial condition. In this paper we consider the combination of a specification in the unobserved component format, see Harvey (2005), with marginal likelihood based inference. The model is

$$y_t = \mu + x_t'\beta + u_t, \tag{2.1}$$

$$u_{t+1} = \rho u_t + v_t,\ t = 1, \ldots, n, \tag{2.2}$$

$$u_1 \quad \begin{cases} = \xi & \text{for } \rho = 1, \\ \sim N(0, \sigma_v^2/(1-\rho^2)) & \text{for } |\rho| < 1, \end{cases} \tag{2.3}$$

where v_t is a potentially serially correlated stationary process with standard normality assumptions, x_t is a $(k-1) \times 1$ vector, and ξ is an unknown scalar. The specification of the initial condition (2.3) is coherent as the variance of u_1 goes to infinity for $\rho \uparrow 1$. The method of

[1] This chapter is based on Francke and de Vos (2006).

estimating and testing is based on the marginal likelihood.

The marginal likelihood is a concept suited for inference on parameters in the covariance matrix of the general linear model. Seminal articles on this approach are Harville (1974), King (1980) and Tunnicliffe Wilson (1989). The $(k \times 1)$ regression parameter vector $(\mu, \beta')'$ and the scale parameter σ are considered as nuisance parameters. A maximal invariant transformation can be found such that an $(n - k - 1)$ dimensional distribution of the transformed data only depends on the parameters of the covariance matrix. The likelihood of this transformation is called the marginal likelihood.

For model (2.1)–(2.3) an essential part of the transformation is based on first differences, leading to $\Delta y_t = \Delta x_t'\beta + \Delta u_t$ and $\Delta u_{t+1} = (\rho - 1)u_t + v_t$. This transformation reduces the dimension to $(n - 1)$ and removes the nuisance parameters μ, the equilibrium level, as well as the initial condition ξ. The variance of Δu_{t+1} is provided by $\mathrm{Var}(\Delta u_{t+1}) = 2/(1 + \rho)\sigma_v^2$, so Δu_{t+1} is stationary for $-1 < \rho \leq 1$ and the marginal likelihood is well behaved when $\rho = 1$: finite, nonzero and continuous for $\rho \uparrow 1$. Unlike in the full likelihood there is no "unit root problem" in the marginal likelihood.

The autocorrelation parameter ρ can be estimated and used for testing the null hypothesis of $\rho = 1$, by marginal likelihood ratio (MLR) and related tests.

In the standard first-order autoregressive model (AR(1)) with constant or with constant and trend, where v_t are i.i.d. Gaussian with variance σ^2, the power of the tests appears to be very close to the theoretical upper bound. The results appear to be better than the tests developed by Fuller (1976) and Dickey and Fuller (1979), Elliott, Rothenberg, and Stock (1996) and—in small samples—Elliott (1999). Moreover the asymptotic approximations in terms of the local-to-unity representation $\gamma = n(1 - \rho)$ are accurate in samples as small as $n = 25$.

When v_t has serial correlation, model selection and generalized MLR tests may be used. We investigated the case that v_t is a first-order moving average (MA(1)) process. The tests for $\rho = 1$ show a reduction in power, but with a very small size distortion. In case of unknown serial correlation in v_t asymptotic results may be applied. We derive the asymptotic distribution of the MLR and the score function in terms of γ. These distributions depend on the correlation structure of v_t only by the ratio κ, the unconditional variance of v_t divided by its long term variance. Conditional on κ the adjusted MLR and score function can be used for estimation and testing. The ratio κ can be estimated consistently. The results are promising, but depend strongly on the quality of the estimation procedure for κ.

The marginal likelihood is invariant with respect to μ, β, and σ^2. Tests that are closely related to the MLR tests are those given by Elliott (1999) and Müller and Elliott (2003), henceforth ME. They use the same specification under various assumptions on ξ and invariance arguments, but they do not derive a likelihood; their unit root tests are pointwise optimal, like those of Dufour and King (1991).

Another closely related approach is followed by Pere (2003). He derives the adjusted profile likelihood in AR(1) models. As this is, apart from an (in this case irrelevant) proportionality

constant $|X'X|$, equal to the marginal likelihood, he would have obtained the same results if he had used the same specification. However, he treats the initial observation differently and obtains different results. As the adjusted profile likelihood is not defined as a proper likelihood of reduced dimension, it is less clear than in the case of the marginal likelihood how to treat the initial observation.

Other likelihood based approaches of unit root testing can also be found in literature. Pantula, Gonzalez-Farias, and Fuller (1994) consider tests based on unconditional full maximal likelihood estimation. Some of their simulation results are close to the ones we will present. However, their tests are not likelihood ratio tests, because the full likelihood is zero for $\rho = 1$. Shin and Fuller (1998) use similar inference for the model $y_t = \rho y_{t-1} + \varepsilon_t$. Lee and Dickey (2004) extend this to the seasonal model $y_t = \rho y_{t-d} + \varepsilon_t$. Their results cannot be compared to ours, because they use a different specification. In these models the level of the series is unknown for $\rho = 1$ and the equilibrium level is known to be zero for $\rho < 1$. This leads to high power of their tests compared to specification (2.1)–(2.3) where the equilibrium level is assumed unknown for $\rho < 1$. Like Bhargava (1986), Schmidt and Phillips (1992), and Harvey (2005) we argue that model (2.1)–(2.3) is a realistic specification because it has a coherent meaning for the expected level of the process $(\mu + x_t'\beta)$ with $|\rho| < 1$.

Marginal likelihood has an additional advantage if there are explanatory variables x_t. Inference on (μ, β, σ^2) is made independent of inference on the covariance matrix. In general this leads to powerful tests, as King (1980) and Rahman and King (1997) show in several examples.

The setup of this chapter is as follows. In section 2.2 we give a simple derivation of the classical marginal likelihood in the general linear model. Next the marginal likelihood for the linear model with AR(1) disturbances is derived. Section 2.3 concerns two basic models where v_t are i.i.d. Gaussian with variance σ^2: AR(1) with constant, and AR(1) with constant and trend. First the power envelope in finite samples is derived, which is possible as the marginal likelihood contains only one parameter. Next asymptotic distributions for the MLR and score function are derived. Section 2.3.4 presents simulation results. MLR tests and tests based on estimates of the AR parameter ρ appear to perform equally well. It is explained why the power functions are close to the theoretical upper bounds. Section 2.4 concerns situations where the v_t are serially correlated. Test results are provided for the case v_t follows a MA(1) process and for a procedure based on the estimation of κ. Section 2.5 concludes.

2.2 Classical marginal likelihood

2.2.1 Marginal likelihood in the general linear model

In this section we give a simple derivation of the marginal likelihood in the general linear model $y = X\beta + u$, with $u \sim N(0, \sigma^2\Omega)$, Ω a positive definite matrix depending on a n_θ dimensional vector θ, so $\Omega = \Omega(\theta)$, and X an $(n \times k)$ matrix of regressors with rank k. We are interested in

inference on θ, and regard β (and later on also σ^2) as nuisance parameters. The full likelihood for this model is

$$f(y|\theta,\sigma^2,\beta) = (2\pi\sigma^2)^{-n/2}\left|\Omega\right|^{-1/2}\exp\left\{-\frac{y'\Omega^{-1}M_X^{\Omega}y}{2\sigma^2}\right\}, \tag{2.4}$$

with $M_X^{\Omega} = I - X(X'\Omega^{-1}X)^{-1}X'\Omega^{-1}$.

The concept of marginal likelihood was introduced by Kalbfleisch and Sprott (1970). For the linear model it is used in the context of unbalanced incomplete block designs by Patterson and Thompson (1971), who refer to it as the likelihood of error contrasts. The use of the classical marginal likelihood is limited to location and scale parameters and some other applications, which may explain that it remained relatively unknown.

The idea of marginal likelihood (and error contrasts) in the general linear model is that the likelihood (2.4) can be expressed as the product of two likelihoods associated with transformations $A'y$ and $B'y$, so

$$f(y|\theta,\sigma^2,\beta) = f(A'y|\theta,\sigma^2)\times f(B'y|\theta,\sigma^2,\beta), \tag{2.5}$$

where A is an $n\times m$ matrix, $m = n-k$, and B an $n\times k$ matrix.

The likelihood of the transformed data $A'y$ is called the marginal likelihood (with respect to β), if the matrix A meets the following conditions: $A'X = 0$ and $r(A) = m$. The marginal likelihood $L_{M_\beta}(\theta,\sigma^2)$ is the density function $f(A'y|\theta,\sigma^2)$ and is provided by

$$L_{M_\beta}(\theta,\sigma^2) = (2\pi\sigma^2)^{-m/2}\left(\frac{|X'X|}{|A'A|\,|X'\Omega^{-1}X|\,|\Omega|}\right)^{1/2}\exp\left\{-\frac{y'\Omega^{-1}M_X^{\Omega}y}{2\sigma^2}\right\}. \tag{2.6}$$

The data occur in the likelihood only in the residual sum of squares $y'\Omega^{-1}M_X^{\Omega}y$. Because $\Omega^{-1}M_X^{\Omega}$ has rank m the marginal likelihood actually corresponds to an m dimensional density. The choice of A is irrelevant for inference on θ provided A does not depend on θ. The most elegant choice is $A'A = I_m$, a unique choice up to an orthonormal transformation of columns, as introduced by Harville (1974). An equivalent notation then is $f(M_Xy|\theta,\sigma^2)$, so the marginal likelihood can be defined as the density of the ordinary least squares residuals, but one must keep in mind that this distribution is degenerate (M_Xy is a vector of size n with an m dimensional likelihood). From now on it is assumed that $A'A = I_m$. In this case the marginal likelihood is invariant to regular transformations of X.

It follows from (2.5) and (2.6) that

$$f(B'y|\theta,\sigma^2,\beta) = (2\pi\sigma^2)^{-k/2}\left(\frac{|X'\Omega^{-1}X|}{|X'X|}\right)^{1/2}\exp\left\{-\frac{(\beta-b)'\,(X'\Omega^{-1}X)\,(\beta-b)}{2\sigma^2}\right\}, \tag{2.7}$$

where $b = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y$. The claim is that $f(B'y|\theta,\sigma^2,\beta)$ contains no available infor-

mation in absence of information on β. According to Patterson and Thompson (1971) "$B'X$ is of rank k with linear functions of $B'y$ exactly determining linear functions of elements of β, leaving no degrees of freedom for further parameter estimation." There appears to be no loss of information on θ and σ^2 by using $A'y$ in place of y, though it is difficult to give a totally satisfactory justification of this claim, see McCullagh and Nelder (1989). Exactly this point has been criticized by Bernardo and Smith (1994, p. 481), who state that for the notion of a "function not containing relevant information in the absence of knowledge about the nuisance parameters", no operational definition has ever been provided.

Various derivations of the marginal likelihood can be found in literature. Harville (1974) gives a derivation based on a decomposition of the likelihood function. More recently, Smyth and Verbyla (1996) give a derivation of what they call "residual likelihood" by conditioning on the sufficient statistics; they apply it also to more complex models. Rahman and King (1997) provide more references as well as a survey of applications.

A direct derivation of the marginal likelihood that we did not find in the literature follows from the observation that

- $A(A'A)^{-1}A' = M_X = I_n - X(X'X)^{-1}X'$,
- $A(A'\Omega A)^{-1}A' = \Omega^{-1}M_X^{\Omega}$, and
- $|A'\Omega A| = |A'A|\,|X'X|^{-1}\,|X'\Omega^{-1}X|\,|\Omega|$,

for any A with $r(A) = m$ and $A'X = 0$. A proof is provided in appendix A.1.

The differences with the profile likelihood (the full likelihood (2.4) maximized with respect to β) are the term $|X'X|^{1/2}\,|X'\Omega^{-1}X|^{-1/2}$ and the power of $\sigma : m$ instead of n. Consequently inference on σ and θ differs. Estimation based on marginal likelihood is to be preferred since it adjusts for the evidence on θ in the part of the data that is a linear function of X, which is pseudo-information. The case with k observations is insightful since $\Omega = \Omega(\theta)$ drops out of the marginal likelihood. Shephard (1993) and Kuo (1999) show that the score function of the marginal likelihood has zero expectation as the marginal likelihood is based on the density of a random variable. Therefore the maximum marginal likelihood can give unbiased estimates.

For inference on θ, the scale parameter σ is still a nuisance parameter. It may be removed by any transformation to ratios of elements of $A'y$. The generally applicable transformation is $y^* = A'y/\sqrt{y'AA'y}$, leading to

$$L_{M_{\beta,\sigma}}(\theta) := f(y^*|\theta) = \frac{\frac{1}{2}\Gamma(\frac{m}{2})|X'X|^{1/2}}{\pi^{m/2}\,|X'\Omega^{-1}X|^{1/2}\,|\Omega|^{1/2}}\left(\frac{y'\Omega^{-1}M_X^{\Omega}y}{y'M_Xy}\right)^{-m/2}, \tag{2.8}$$

an $(m-1)$ dimensional marginal likelihood, independent of β and σ^2, see King (1980). He shows that y^* is a maximal invariant under the group of transformations

$$y \to \eta_0 y + X\eta, \tag{2.9}$$

where η_0 is a positive scalar and η is a $k \times 1$ vector. The principle of invariance implies that we can treat the maximal invariant y^* as the observed random vector and (2.8) as its density function, and therefore as a likelihood function for θ, see Rahman and King (1997).

The marginal likelihood $L_{M_{\beta,\sigma}}(\theta)$ is proportional to the concentrated marginal likelihood $L_{M_\beta}(\theta, \sigma^2_{\mathrm{ML}})$, which may be computed by substituting $\widehat{\sigma}^2_{\mathrm{ML}}(\theta)$ in (2.6), where $\widehat{\sigma}^2_{\mathrm{ML}}(\theta) = \arg\max \ell_{M_\beta}(\theta, \sigma^2) = y'\Omega^{-1}M^{\Omega}_X y/m$, $\ell_{M_\beta}(\theta, \sigma^2) = \ln L_{M_\beta}(\theta, \sigma^2)$. For most applications this means that concentration and marginalization of σ^2 are equivalent. In this article the marginal likelihood is used for testing the hypothesis $\rho = 1$ in first-order autoregressive models. Often we will use $L_{M_\beta}(\rho, \sigma^2)$ in derivations to concentrate with respect to σ^2 later on. $L_{M_{\beta,\sigma}}(\rho)$ is mainly important to analyze the theoretical properties of the tests.

Some further historical notes are interesting. Levenbach (1972) derives the marginal likelihood for first and second order autoregressive models without explanatory variables in a different manner, but did not use the marginal likelihood for unit root testing. Cooper and Thompson (1977) use the marginal likelihood to estimate the parameters in an ARIMA model.

As noted by Tunnicliffe Wilson (1989) in the linear model the modified and the adjusted profile likelihood are proportional to the marginal likelihood. The adjusted profile likelihood from Cox and Reid (1987, 1993) is a simplification of the modified profile likelihood from Barndorff-Nielsen (1983). In the linear model both likelihoods coincide. The motivation for these methods is quite different from that of the marginal likelihood. They are adjustments of the profile likelihood based on asymptotically sufficient statistics for β and σ^2.

2.2.2 Marginal likelihood for linear models with AR(1) disturbances

Consider the first-order autoregressive model as specified in (2.1)–(2.3) without explanatory variables x_t, where the $v_t = \varepsilon_t \sim N(0, \sigma^2)$ are independent, giving

$$y_t = \mu + u_t, \tag{2.10}$$

$$u_{t+1} = \rho u_t + \varepsilon_t, \tag{2.11}$$

$$u_1 \quad \begin{cases} = \xi & \text{for } \rho = 1. \\ \sim N(0, \sigma^2/(1-\rho^2)) & \text{for } |\rho| < 1. \end{cases} \tag{2.12}$$

For $|\rho| < 1$, this is a stationary autoregressive model with unknown level μ, and $u_1 \sim N(0, \sigma^2/(1-\rho^2))$, so the marginal likelihood (with respect to μ) is an $(n-1)$ dimensional likelihood. When $\rho = 1$, it represents a random walk without drift, $y_t = y_{t-1} + \varepsilon_{t-1}$, for $t = 2, \ldots, n$. In both cases the marginal likelihood is an $(n-1)$ dimensional likelihood.

The fact that the marginal likelihood exists, even for $\rho = 1$, can simply be seen by expressing the model in first differences of the data: $y_t - y_{t-1} = (\rho - 1)u_{t-1} + \varepsilon_{t-1}$. The unconditional variance is given by $\mathrm{Var}(y_t - y_{t-1}) = 2\sigma^2/(1+\rho)$, and so the marginal likelihood is well-defined for $-1 < \rho \leq 1$.

For $\rho = 1$, the marginal loglikelihood ℓ_{M_β} is given by

$$-2\ell_{M_\beta}(\rho = 1, \sigma^2) = (n-1)\ln 2\pi\sigma^2 + \sigma^{-2}\Sigma_{t=2}^{n}(y_t - y_{t-1})^2 - \ln n. \tag{2.13}$$

Note that this is the likelihood of the first differences minus a term $\ln n$. The latter is the Jacobian of the transformation from the first differences to $A'y$.

From (2.6) it follows that the marginal loglikelihood $\ell_{M_\beta}(\rho, \sigma^2)$ for (2.10)–(2.12) is given by

$$-2\ell_{M_\beta}(\rho, \sigma^2) = (n-1)\ln 2\pi\sigma^2 + \ln\left|\mathbf{i}'\Omega^{-1}\mathbf{i}\right| + \ln|\Omega| - \ln|n| + \sigma^{-2}\mathrm{RSS}_\mu(\rho), \tag{2.14}$$

where

$$\mathrm{RSS}_\mu(\rho) = (1-\rho^2)y_1^2 + \sum_{t=2}^{n}(y_t - \rho y_{t-1})^2 - \frac{1-\rho}{n-(n-2)\rho}\left(y_1 + (1-\rho)\sum_{t=2}^{n-1}y_t + y_n\right)^2 \tag{2.15}$$

is the generalized least squares residual sum of squares conditional on ρ, $\mathbf{i}$ is an $(n \times 1)$ vector of ones, and the element ij of the variance matrix Ω is given by $\Omega_{ij} = \rho^{|i-j|}/(1-\rho^2)$. As $|\Omega| = (1-\rho^2)^{-1}$, and $|\mathbf{i}'\Omega^{-1}\mathbf{i}| = (n-(n-2)\rho)(1-\rho)$ for $\rho \neq 1$ a factor $(1-\rho)$ cancels, and for $-1 < \rho < 1$ the marginal likelihood can be expressed as

$$-2\ell_{M_\beta}(\rho, \sigma^2) = (n-1)\ln 2\pi\sigma^2 + \ln\frac{n-(n-2)\rho}{n(1+\rho)} + \sigma^{-2}\mathrm{RSS}_\mu(\rho). \tag{2.16}$$

For equation (2.14) the limit for $\rho \uparrow 1$ exists and is provided by (2.16) with $\rho = 1$.

The profile loglikelihood ℓ_P, given by

$$-2\ell_P(\rho, \sigma^2) = (n-1)\ln 2\pi\sigma^2 - \ln(1-\rho^2) + \sigma^{-2}\mathrm{RSS}_\mu(\rho),$$

is not defined when $\rho = 1$, due to the term $\ln(1-\rho^2)$.

In the model (2.1)–(2.3) with explanatory variables x_t where the $v_t = \varepsilon_t$ are independent, for $|\rho| < 1$, the marginal likelihood (with respect to μ and β) is m dimensional. When $\rho = 1$, it is a specification in first differences, $y_t - y_{t-1} = (x_t - x_{t-1})\beta + \varepsilon_{t-1}$, for $t = 2, \ldots, n$; after marginalization with respect to β the marginal likelihood is also m dimensional.

If we define the following,

$$\begin{array}{lcllcl}
y_t(\rho) & = & y_t - \rho y_{t-1}, & \Sigma_{yy,\rho} & = & (1-\rho^2)y_1^2 + \Sigma_{t=2}^{n} y_t(\rho)^2, \\
x_t(\rho) & = & x_t - \rho x_{t-1}, & \Sigma_{xy,\rho} & = & (1-\rho^2)x_1'y_1 + \Sigma_{t=2}^{n} x_t(\rho)'y_t(\rho), \\
\Sigma_{y,\rho} & = & y_1 + (1-\rho)\Sigma_{t=2}^{n-1} y_t + y_n, & \Sigma_{xx,\rho} & = & (1-\rho^2)x_1'x_1 + \Sigma_{t=2}^{n} x_t(\rho)'x_t(\rho), \\
\Sigma_{x,\rho} & = & x_1 + (1-\rho)\Sigma_{t=2}^{n-1} x_t + x_n, & F^{-1} & = & g(\rho) - (1-\rho)\Sigma_{x,\rho}\Sigma_{xx,\rho}^{-1}\Sigma_{x,\rho}', \\
\widehat{\mu} & = & F\left(\Sigma_{y,\rho} - \Sigma_{x,\rho}\Sigma_{xx,\rho}^{-1}\Sigma_{xy,\rho}\right), & \widehat{\beta} & = & \Sigma_{xx,\rho}^{-1}\Sigma_{xy,\rho} - (1-\rho)\Sigma_{xx,\rho}^{-1}\Sigma_{x,\rho}'\widehat{\mu}, \\
\tilde{X} & = & (I_n - \frac{1}{n}\mathbf{i}\mathbf{i}')X, & g(\rho) & = & n - (n-2)\rho,
\end{array}$$

then the residual sum of squares if provided by

$$\mathrm{RSS}_{\mu,\beta}(\rho) = \Sigma_{yy,\rho} - (1-\rho)\Sigma_{y,\rho}\widehat{\mu} - \Sigma'_{xy,\rho}\widehat{\beta}, \tag{2.17}$$

and for $|\rho| < 1$ the marginal loglikelihood $\ell_{M_\beta}(\rho, \sigma^2)$ can be expressed as

$$\begin{aligned} -2\ell_{M_\beta}(\rho,\sigma^2) = m\ln 2\pi\sigma^2 + \ln\frac{g(\rho)}{n(1+\rho)} - \ln|\tilde{X}'\tilde{X}| \\ + \ln|\Sigma_{xx,\rho} - \frac{1-\rho}{g(\rho)}\Sigma'_{x,\rho}\Sigma_{x,\rho}| + \sigma^{-2}\mathrm{RSS}_{\mu,\beta}(\rho). \end{aligned} \tag{2.18}$$

The existence of the marginal likelihood in $\rho = 1$ follows from the fact that the marginal likelihood of $A'y$ is proportional to the likelihood of $B'D'y$, where D is the matrix of first differences and B a matrix of full column rank such that $B'(D'X) = 0$. Now $B'D'y \sim N(0, B'D'\Omega DB)$, where Ω is the covariance matrix of the AR(1) process, and $D'\Omega D$ is well defined for $-1 < \rho \leq 1$. The difference between the marginal likelihood and the likelihood of $B'D'y$ is a Jacobian term $\ln|\Sigma_{xx,1}| - \ln|\tilde{X}'\tilde{X}| - \ln n$. Note that for $\rho = 1$ this transformation implies that the loglikelihood of $B'D'y$ equals $m\ln 2\pi\sigma^2 + \sigma^{-2}\mathrm{RSS}_{\mu,\beta}(1)$.

Substitution of $\widehat{\sigma}^2_{\mathrm{ML}} = \mathrm{RSS}_{\mu,\beta}(\rho)/m$ in (2.18) gives $\ell_{M_\beta}(\rho, \widehat{\sigma}^2_{\mathrm{ML}}) \propto \ell_{M_{\beta,\sigma}}(\rho)$, the marginal likelihood for ρ. Again, unlike the profile likelihood, this marginal likelihood is also defined for and continuous in $\rho = 1$, and suited as a basis for unit root tests.

2.3 Unit root tests in some basic models

2.3.1 Marginal likelihood tests

Many unit root tests are described in the econometric literature for the model with a constant, and for the model with a constant and linear trend. The classics are those developed in Fuller (1976) and Dickey and Fuller (1979). More recent tests are e.g. the P_T test by Elliott, Rothenberg, and Stock (1996), the Q_T tests by Elliott (1999). We propose two new ones, based on marginal likelihood. Define $\widehat{\rho}_{\mathrm{ML}} = \arg\max \ell_{M_{\beta,\sigma}}(\rho)$, the maximum marginal likelihood estimator of ρ. The first test is a marginal loglikelihood difference test, evaluated in $\widehat{\rho}_{\mathrm{ML}}$, and is given by $T_1 = \ell_{M_{\beta,\sigma}}(\widehat{\rho}_{\mathrm{ML}}) - \ell_{M_{\beta,\sigma}}(1)$. The second test is $T_2 = \widehat{\gamma}_{\mathrm{ML}} = n(1-\widehat{\rho}_{\mathrm{ML}})$, where the local-to-unit root format is used in order to have a useful asymptotic framework to analyse power. In both tests the null hypothesis of a unit root is rejected if the test statistic exceeds the critical value, depending on the size α and the number of observations n.

Test T_1 is based on the marginal likelihood ratio. In terms of $\gamma = n(1-\rho)$ the MLR for the model with constant (superscript μ) and the model with constant and trend (superscript τ) is

provided by

$$\mathrm{MLR}^i(\gamma) = h^i(\gamma)^{1/2} \left(\frac{\mathrm{RSS}_i(\gamma)}{\mathrm{RSS}_i(\gamma = 0)} \right)^{-(n-k_i)/2}, \tag{2.19}$$

where $i = \mu, \tau$, $k_\mu = 1$, $k_\tau = 2$, and $h^i(\gamma)$ are ratios of the determinant terms $|X'X|$, $|X'\Omega^{-1}X|$, and $|\Omega|$ under the null and alternative hypothesis. Test T_1 can equivalently be formulated as $T_1 = \ln \mathrm{MLR}^i(\widehat{\gamma}_{\mathrm{ML}})$.

T_1 is an optimal invariant procedure: it depends on a function of a maximal invariant, see Lehmann (1986). Dufour and King (1991) also developed location and scale maximal invariant tests, but outside the maximum likelihood context and thus only pointwise optimal. The $\overline{Q}(\gamma)$ statistic from ME is only location invariant, but scale invariance appears to be (as they remark on p. 1274) asymptotically irrelevant. The relation with the marginal likelihood ratio is given by

$$\mathrm{MLR}^i(\gamma) = h^i(\gamma)^{1/2} \left(1 + \frac{\overline{Q}^i(\gamma)}{\mathrm{RSS}_i(\gamma = 0)} \right)^{-(n-k_i)/2}.$$

For fixed γ the resulting tests are asymptotically equivalent. However $\mathrm{MLR}^i(\gamma)$ may be optimized with respect to γ. The resulting test statistic $T_1 = \ln \mathrm{MLR}^i(\widehat{\gamma}_{\mathrm{ML}})$ does not require the choice of γ and may be expected to have better power against alternatives other than the chosen γ.

2.3.2 The power envelope in finite samples

The power envelope can be used as a benchmark for the power function of tests. It is also possible to derive the finite and asymptotic power envelope when the v_t are i.i.d. normal variables. The marginal likelihood depends on only one parameter. The Neyman-Pearson lemma defines the optimal critical region for any fixed alternative $\gamma = n(1 - \rho)$. The power envelope can be computed as

$$P_\gamma \left(\ln \mathrm{MLR}^i(\gamma) > \kappa_\alpha(\gamma, n) \right),$$

where the subscript γ indicates the data generating process and $\kappa_\alpha(\gamma, n)$ is defined such that

$$P_{\gamma=0} \left(\ln \mathrm{MLR}^i(\gamma) > \kappa_\alpha(\gamma, n) \right) = 1 - \alpha.$$

Table 2.1 provides power envelopes for maximal invariant unit root tests in the AR(1) model with constant (and trend), when n = 25, 50, 100, 250, and 1,000. The data is generated by (2.1)–(2.3), where the v_t are independent standard normal variables. Note that the power envelope in the local-to-unity format is almost insensitive to the number of observations.

Table 2.1: Power envelopes for maximal invariant unit root tests in the AR(1) model.

$\gamma = n(1-\rho)$		0	5	10	15	20
AR(1) with constant	$n = 25$	0.050	0.197	0.543	0.867	0.981
	50	0.050	0.197	0.528	0.848	0.976
	100	0.050	0.196	0.521	0.838	0.973
	250	0.050	0.195	0.516	0.833	0.971
	1000	0.050	0.196	0.515	0.829	0.970
AR(1) with constant and trend	$n = 25$	0.050	0.101	0.272	0.581	0.856
	50	0.050	0.106	0.267	0.540	0.810
	100	0.050	0.101	0.256	0.519	0.782
	250	0.050	0.100	0.252	0.502	0.761
	1000	0.050	0.101	0.251	0.500	0.758

2.3.3 Asymptotic distribution of likelihood ratio and score function

The asymptotic distributions of the MLR and score function are derived for the AR(1) model with constant (and trend) in the local-to-unity framework, where the sample n size goes to infinity and $\gamma = n(1-\rho)$ is a fixed constant. The distributions appear to be polynomial expressions in functions of Brownian motions, which implies that the test statistics T_1 and T_2 have nonstandard asymptotic distributions.

Define $A_n = n^{-2}\sum_{t=1}^{n}\widetilde{u}_t^2$, $B_n = n^{-1}\sum_{t=2}^{n} v_t\widetilde{u}_{t-1}$, $C_n = n^{-3/2}\sum_{t=1}^{n}\widetilde{u}_t$, $D_n = n^{-1/2}\widetilde{u}_n$, $E_n = n^{-5/2}\sum_{t=1}^{n} t\widetilde{u}_t$, and $\widetilde{u}_t = u_t - u_1$. u_t is generated by (2.2)–(2.3), where $\rho = \rho_0$, so $\gamma_0 = n(1-\rho_0)$.

Assume that v_t has a moving average presentation $v_t = \sum_{j=1}^{\infty}\psi_j\varepsilon_{t-j}$ where ε_t are independent standard normal variables and $\sum_{j=1}^{\infty} j\,|\psi_j| < \infty$. Elliott (1999) shows in Lemma 2 that

$$n^{-1/2}\left(u_{[Ts]} - u_1\right) \Longrightarrow \omega M(s) = \begin{cases} \omega W(s) & \text{for } \gamma_0 = 0, \\ \omega\left(W_{\gamma_0}(s) + \left(\exp\left(-\gamma_0 s\right) - 1\right)\zeta\right) & \text{else,} \end{cases}$$

where $W_{\gamma_0}(s) = \gamma_0\int_0^s e^{-\gamma_0(s-\lambda)}W(\lambda)\,d\lambda + W(s)$ is an Ornstein Uhlenbeck process, $W(s)$ is a standard Brownian motion, $\zeta \sim N(0, (2\gamma_0)^{-1})$ and is independent of $W_{\gamma_0}(s)$, [.] indicates the greatest lesser integer function, and $\omega^2 = \sum_{j=-\infty}^{\infty}\gamma_v(j)$, and $\gamma_v(j) = \mathrm{E}\left[v_t v_{t-j}\right]$.

It follows that

$$\begin{aligned} A_n &\Rightarrow \omega^2 A_\infty = \omega^2\int_0^1 M(s)^2 ds, \\ C_n &\Rightarrow \omega C_\infty = \omega\int_0^1 M(s) ds, \\ D_n &\Rightarrow \omega D_\infty = \omega M(1), \\ E_n &\Rightarrow \omega E_\infty = \omega\int_0^1 sM(s) ds, \end{aligned}$$

and

$$\widehat{\sigma}_n^2 = D_n^2 - 2B_n + 2\gamma A_n - 2n^{-1}\gamma D_n^2 \to \gamma_v(0) = \mathrm{E}\left[v_t^2\right].$$

Substitution of these results into (2.19) provides the asymptotic distribution of the MLR function under the null hypothesis and local-to-unity alternatives,

$$f_{\gamma_0}\left(\mathrm{MLR}_\infty^i(\gamma)\right) \sim h_\infty^i(\gamma)^{1/2} \exp\left(-\frac{1}{2}\left(\kappa\left[\gamma^2 A_\infty + \gamma D_\infty^2 - R_\infty^i\right] - \gamma\right)\right), \tag{2.20}$$

for $i = \mu, \tau$, where the subscript γ_0 indicates the data generating process and

$$\begin{array}{rclrcl} h_\infty^\mu(\gamma) & = & \frac{2}{\gamma+2}, & R_\infty^\mu & = & \frac{\gamma}{\gamma+2}\left(\gamma C_\infty + D_\infty\right)^2, \\ h_\infty^\tau(\gamma) & = & \frac{24}{(\gamma+2)(\gamma^2+6\gamma+12)}, & R_\infty^\tau & = & \frac{\left(\gamma^2 E_\infty + (\gamma+1)D_\infty\right)^2}{(\gamma^2/3+\gamma+1)} + K_\infty(\gamma), \end{array} \tag{2.21}$$

and

$$K_\infty(\gamma) = \frac{\gamma}{\gamma + 2 - \gamma\frac{(1+\gamma/2)^2}{(\gamma^2/3+\gamma+1)}}\left(\gamma C_\infty + D_\infty - \frac{(1+\gamma/2)}{(\gamma^2/3+\gamma+1)}\left(\gamma^2 E_\infty + (\gamma+1)D_\infty\right)\right)^2.$$

In this section we only consider the case that $\kappa = 1$, where $\kappa = \omega^2/\gamma_v(0)$. The asymptotic distribution of the MLR is independent of μ, β and σ^2, and is a linear combination of more than one statistic, with weights that depend on γ. It follows that even asymptotically no uniformly most powerful (UMP) test exists.

The asymptotic distribution of T_2, the maximum marginal likelihood estimator $\widehat{\gamma}_{\mathrm{ML}} = \arg\max \mathrm{MLR}(\gamma)$, is a function of A_∞, C_∞, D_∞, and E_∞ (for the trend case). No closed form solution is available for the maximization problem. The asymptotic distribution of T_1 can be derived from T_2 by substituting $\widehat{\gamma}_{\mathrm{ML}}$ in (2.20). It is also a function of A_∞, C_∞, D_∞, and E_∞, but it is not a monotone transformation of T_2. So differences in power functions between T_1 and T_2 may be expected.

The maximization of the MLR may partly be done analytically by computing the score function, though this is only feasible for the AR(1) model with constant. The distribution of the asymptotic score function $S_\infty^\mu(\gamma)$ is then given by

$$f_{\gamma_0}\left(S_\infty^\mu(\gamma)\right) \sim -\gamma\left(A_\infty - C_\infty^2\right) - \frac{(C_\infty - D_\infty)^2 + C_\infty^2 - 1}{2} + \frac{(D_\infty - 2C_\infty)^2}{(\gamma+2)^2} - \frac{1}{2(\gamma+2)}.$$

To allow the zero boundary solution, the optimization problem invokes the Kuhn-Tucker conditions, $\gamma S_\infty^i(\gamma) = 0$, and $S_\infty^i(\gamma) + \gamma S_\infty^{i\prime}(\gamma) \leq 0$, where $S_\infty^{i\prime}(\gamma)$, the first derivative of the score function. The first derivative (and higher order) of the asymptotic score function is random,

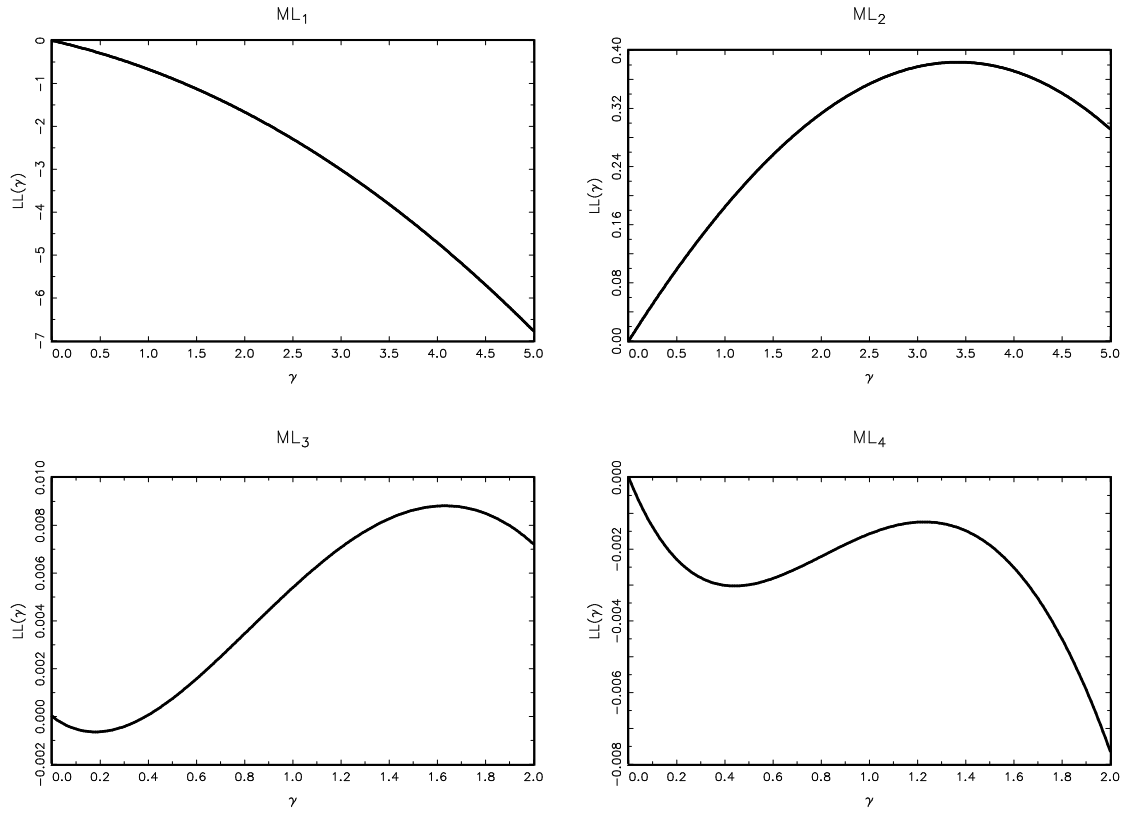

Figure 2.1: Possible forms of the marginal likelihood.

so standard asymptotic distribution theory does not apply. The asymptotic distribution of the first derivative of the score function is provided by

$$f_{\gamma_0}(S_\infty^{\mu\prime}(\gamma)) \sim -\left(A_\infty - C_\infty^2\right) - 2\frac{(D_\infty - 2C_\infty)^2}{(\gamma+2)^3} + \frac{1}{2(\gamma+2)^2}.$$

The asymptotic score function $S_\infty^\mu(\gamma)$ is a polynomial of order three divided by $(\gamma+2)^2$, and has 0, 1, or 2 positive roots, which are functions of A_∞, C_∞ and D_∞. The score function for finite samples $S_n^\mu(\gamma)$ is a fourth order polynomial, which can conveniently be used to compute $\widehat{\gamma}_{\mathrm{ML}}$. Some care in optimization is required as indicated by figure 2.1 where the possible forms of the marginal likelihood are shown. ML_3 and ML_4 have two local optima. In our simulations we use the root with highest marginal likelihood. As ML_4 rarely occurs, only 84 times in a simulation study of 300,000, this choice is not relevant.

The asymptotic distribution of T_2 is the limit distribution of $\widehat{\gamma}_{\mathrm{ML}}$ and is continuous except at zero, and has a masspoint in zero, with $P(\widehat{\gamma}_{\mathrm{ML}} = 0) \approx P\left(S_\infty^\mu(0) \leq 0\right) = P\left(\chi^2(1) > 1\right) \approx 0.317$.

Table 2.2: Critical values for unit root test in the AR(1) model with constant.

	$T_1 = \ell_{M_{\beta,\sigma}}(\widehat{\rho}_{\mathrm{ML}}) - \ell_{M_{\beta,\sigma}}(1)$				$T_2 = n(1-\widehat{\rho}_{\mathrm{ML}})$			
$\alpha\backslash n$	50	100	250	1000	50	100	250	1000
0.01	3.189	3.213	3.224	3.236	16.4	17.0	17.3	17.5
0.025	2.375	2.399	2.408	2.420	13.1	13.5	13.7	13.8
0.05	1.784	1.799	1.803	1.811	10.6	10.8	10.9	11.0
0.10	1.208	1.221	1.227	1.229	8.0	8.1	8.1	8.1

Table 2.3: Critical values for unit root test in the AR(1) model with constant and trend.

	$T_1 = \ell_{M_{\beta,\sigma}}(\widehat{\rho}_{\mathrm{ML}}) - \ell_{M_{\beta,\sigma}}(1)$				$T_2 = n(1-\widehat{\rho}_{\mathrm{ML}})$			
$\alpha\backslash n$	50	100	250	1000	50	100	250	1000
0.01	3.102	3.134	3.136	3.143	21.9	23.0	23.6	24.0
0.025	2.284	2.310	2.316	2.321	18.3	19.0	19.4	19.6
0.05	1.692	1.712	1.716	1.719	15.4	15.9	16.2	16.3
0.10	1.120	1.136	1.137	1.141	12.3	12.6	12.7	12.8

2.3.4 Empirical results

In this section power functions for different unit root tests are compared. In a simulation study the data are generated by (2.1)–(2.3), where the v_t are independent standard normal variables. The choice of μ, β and σ^2 is irrelevant, as it follows from section 2.2 that the marginal likelihood is independent of these parameters.

In tables 2.2 and 2.3 critical values for the tests T_1 and T_2 are provided, based on the 90%, 95%, 97.5% and 99% quantiles of the distributions. The entries of the tables are based on 1,000,000 Monte Carlo replications. From these tables it appears that the critical values for both tests (T_1 and T_2) are practically insensitive to the number of observations. This means that we can use asymptotic critical values even for small samples.

As expected the critical values of test T_2 for the model with trend are larger than those of the model without trend. That this is not the case for test T_1 may be surprising, but one has to consider that the marginal likelihood (ratio) of different models cannot directly be compared with each other. The marginal likelihoods have different dimensions and the terms $|X'\Omega^{-1}X|$ and $|X'X|$ are different for both models. Note that it follows from (2.21) that $h^{\mu}_{\infty}(\gamma) > h^{\tau}_{\infty}(\gamma)$ for $\gamma > 0$.

We first compare the power functions of these two tests to those of three other unit root tests: the Dickey-Fuller τ-statistic and z-statistic, see for example Davidson and MacKinnon (1993), and the P_T statistic. We choose these tests as they, like the tests T_1 and T_2, are invariant with respect to the size of the constant, the coefficient of the trend, and the scale. The tests developed in Dickey and Fuller (1981) do not have this property.

Figure 2.2 and 2.3 show the power functions using 5% level tests for both models for $n = 100$. The power functions are based on 10,000 Monte Carlo replications for $\rho = 0.80, 0.81, \ldots, 1$. The

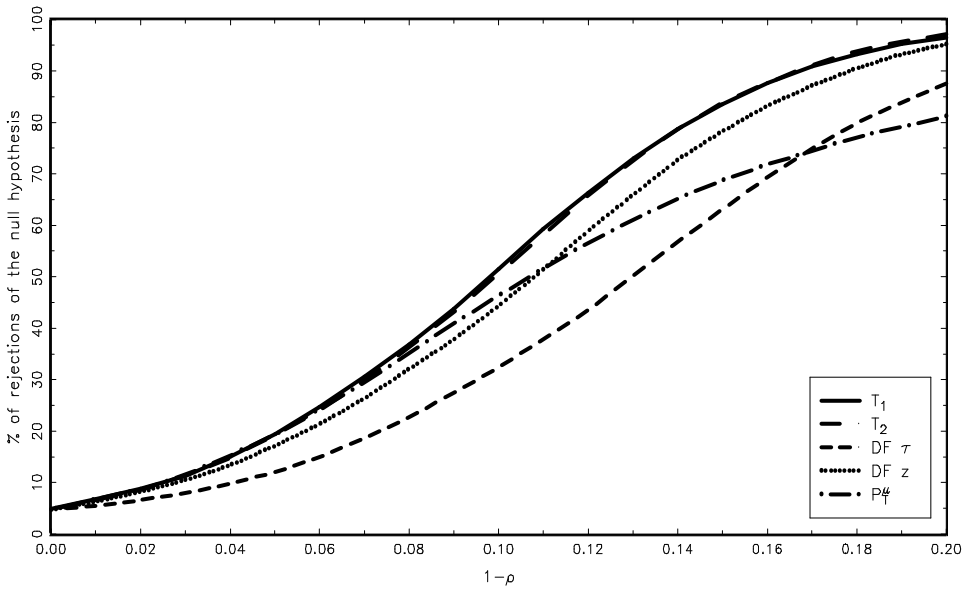

Figure 2.2: Power functions for unit root tests for the model with constant.

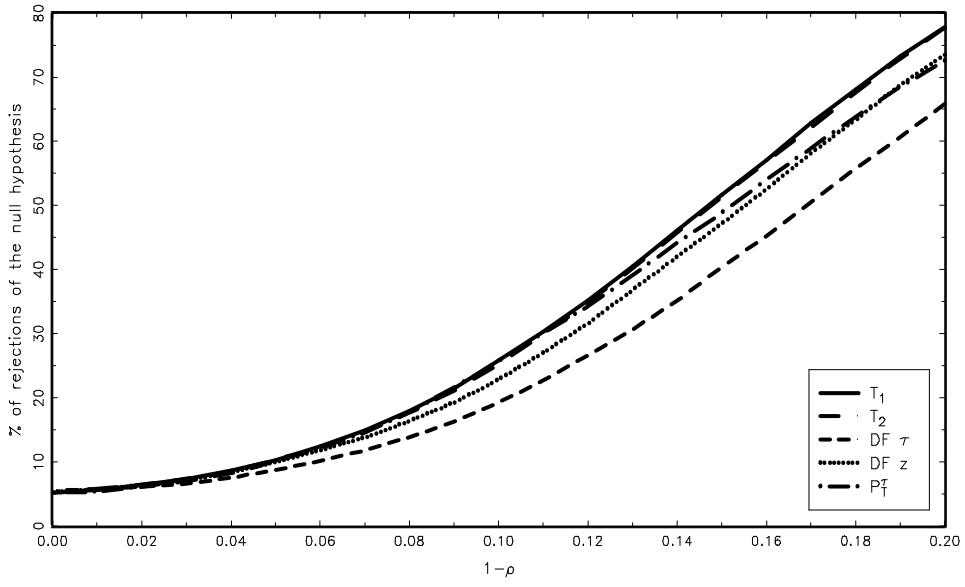

Figure 2.3: Power functions for unit root tests for the model with constant and trend.

critical values for the Dickey-Fuller τ test are -2.89 (-3.45), for the z test $0.863 = 1 - 13.7/100$ ($0.793 = 1 - 20.7/100$), and for the P_T test 3.214 (6.055). Between brackets the critical values for the model with constant and trend are provided. The critical values for P_T differ from the ones provided by Elliott, Rothenberg, and Stock (1996) due to the random initial condition and the use of $\widehat{\sigma}^2_{\mathrm{ML}}$. The marginal likelihood unit root tests T_1 and T_2 appear to do equally well and outperform the other unit root tests.

The poor performance of the P_T test confirms the results of Elliott (1999), who shows that this is due to the fact that the initial observation is considered as fixed. He shows that this even affects the asymptotic power envelope. For the case where the first observation is drawn from its unconditional distribution, he derives the asymptotic power envelope and locally efficient tests. His test statistics $Q_T(\gamma)$ are efficient around γ; $Q_T(10)$[2] appears to have best overall performance. We compare the power functions of the marginal likelihood tests T_1 and T_2 to the asymptotic power envelope and with $Q_T(10)$, for $n = 25,\ 50, 100,\ 250$, and 1,000 and $\gamma = 0,\ 5,\ 10,\ 15$, and 20.

Tables 2.4 and 2.5 provide power functions for the marginal likelihood and the $Q_T(10)$ tests based on 10,000 replications. The critical values for $n = 1{,}000$ are used, as approximations of the asymptotic critical values. In the tables this is denoted by T^∞. The results based on T^∞ are close to the results based on the critical values for a specific n, indicated by T^n. The power function for the $Q_T(10)$ test is calculated under the assumption that it is known that the v_t are independent normal variables.

In both cases the MLR test T_1 has no size distortion. One may conclude that the tests T_1 and T_2 are close to their theoretical upper bound as provided in table 2.1, even in small samples ($n = 25$), which is remarkable as no UMP test exist. For large samples the power functions of T_1 and T_2 are very close to the asymptotic power envelope. This is also the case for $Q_T(10)$, however for small values of n there is size distortion for the model with constant and trend. Size adjusted power functions of $Q_T(10)$ are not provided, but are close to the upper bound as provided in table 2.1, like the tests T_1 and T_2.

2.3.5 Why the tests are "almost UMP"

We showed that, in some simple models, tests based on maximum marginal likelihood are remarkably close to the UMPI (uniformly most powerful invariant) upper bound. In this section we give an explanation of this fact for the AR(1) model with constant. Other models with AR(1) disturbances may be expected to have similar characteristics.

Only one parameter remains in the marginal likelihood. If the likelihood ratio is monotone in some statistic, tests based on this statistic are UMPI, see Lehmann (1986). In our case no such statistic exists, but a stochastic variation on the monotone likelihood ratio theorem appears to explain the observed phenomenon.

[2]Elliott's notation is $Q_T(-10)$, he uses the definition $\gamma = n(\rho - 1)$.

Table 2.4: Power functions for unit root test in the AR(1) model with constant.

$\gamma = n(1-\rho)$		0	5	10	15	20
Asymptotic Power Envelope		0.05	0.20	0.52	0.83	0.97
T_1^n	$n = 25$	0.049	0.192	0.541	0.862	0.977
	50	0.049	0.194	0.526	0.845	0.972
	100	0.050	0.192	0.518	0.836	0.969
	250	0.050	0.194	0.513	0.828	0.967
	1000	0.049	0.193	0.513	0.826	0.966
T_2^n	$n = 25$	0.050	0.187	0.531	0.864	0.981
	50	0.049	0.190	0.520	0.846	0.975
	100	0.049	0.189	0.512	0.836	0.972
	250	0.050	0.190	0.506	0.827	0.970
	1000	0.049	0.190	0.506	0.823	0.970
T_1^∞	$n = 25$	0.046	0.181	0.522	0.851	0.975
	50	0.052	0.194	0.523	0.837	0.970
	100	0.048	0.187	0.516	0.836	0.970
	250	0.049	0.201	0.516	0.830	0.967
T_2^∞	$n = 25$	0.039	0.154	0.470	0.824	0.972
	50	0.048	0.179	0.491	0.820	0.968
	100	0.048	0.179	0.499	0.828	0.970
	250	0.049	0.197	0.505	0.824	0.971
$Q_T(10)$	$n = 25$	0.041	0.163	0.489	0.823	0.964
	50	0.044	0.179	0.503	0.826	0.964
	100	0.052	0.192	0.511	0.824	0.962
	250	0.049	0.193	0.522	0.831	0.967
	1000	0.048	0.194	0.518	0.829	0.963

Table 2.5: Power functions for unit root test in the AR(1) model with constant and trend.

$\gamma = n(1-\rho)$		0	5	10	15	20
Asymptotic Power Envelope		0.05	0.10	0.25	0.50	0.76
T_1^n	$n = 25$	0.050	0.097	0.268	0.580	0.857
	50	0.049	0.099	0.258	0.536	0.814
	100	0.050	0.103	0.261	0.521	0.779
	250	0.043	0.096	0.244	0.503	0.766
	1000	0.049	0.096	0.252	0.501	0.757
T_2^n	$n = 25$	0.050	0.094	0.263	0.572	0.863
	50	0.050	0.098	0.254	0.530	0.809
	100	0.050	0.101	0.256	0.517	0.778
	250	0.044	0.092	0.240	0.495	0.762
	1000	0.050	0.096	0.248	0.496	0.755
T_1^∞	$n = 25$	0.048	0.090	0.257	0.567	0.851
	50	0.047	0.096	0.251	0.526	0.806
	100	0.050	0.102	0.259	0.519	0.777
	250	0.043	0.096	0.243	0.503	0.765
T_2^∞	$n = 25$	0.030	0.059	0.178	0.448	0.771
	50	0.041	0.080	0.217	0.473	0.763
	100	0.045	0.093	0.234	0.494	0.760
	250	0.043	0.091	0.236	0.490	0.758
$Q_T(10)$	$n = 25$	0.023	0.049	0.153	0.396	0.697
	50	0.034	0.072	0.202	0.446	0.725
	100	0.046	0.092	0.230	0.477	0.737
	250	0.043	0.091	0.242	0.496	0.754
	1000	0.048	0.101	0.251	0.505	0.757

A graphical illustration explains this. The asymptotic distribution of the MLR, derived in the section 2.3.3, enables us to draw a sample from the asymptotic distribution of the MLR under the null hypothesis, thus when data is generated by a random walk. For any fixed value of γ, the values in this simulation may be ordered according to their MLR(γ). According to the Neyman-Pearson lemma, the 5% points with the lowest ratio constitute the optimal critical region for testing this value of γ. If the ordering is the same for all γ the monotone likelihood ratio theorem applies. This is not the case here; each γ implies a different ordering, but for each $\widehat{\gamma}_{\mathrm{ML}}$ relevant for tests of reasonable size, the test statistics (T_1 and T_2) appear to be monotone functions of the likelihood ratio plus a stochastic variable with a relatively small variance.

For the 5% critical value of the test T_2, $\gamma = 11$, figure 2.4 shows the performance of the MLR test T_1. The values at the right of the vertical bar form the optimal critical zone to test $\gamma = 0$ against $\gamma = 11$. The values above the horizontal bar form the critical zone for the MLR test ($T_1 < 1.81$). Both regions virtually coincide. For $\gamma = 20$ we obtain figure 2.5. The difference between the critical zones is now that, instead of the points left above the crossing which would be optimal for $\gamma = 20$, the points right below the crossing are used. But as the number of points is small (15 out of 10,000 in our sample) and the difference in discriminatory power (represented by the order according to MLR(20)) is small, the performance of the MLR test is very close to the UMPI upper bound.

Similar pictures result for small samples and for test T_2. The MLR test typically coincides with the Neyman-Pearson region near the critical value, as for the nonzero values of $\widehat{\gamma}_{\mathrm{ML}}$,

$$\mathrm{MLR}(\widehat{\gamma}_{\mathrm{ML}}) \cong \mathrm{MLR}(\gamma) + 0.5(\gamma - \widehat{\gamma}_{\mathrm{ML}})^2 \mathrm{MLR}''(\widehat{\gamma}_{\mathrm{ML}}),$$

with no first-order term as $\mathrm{MLR}'(\widehat{\gamma}_{\mathrm{ML}}) = 0$.

The monotonicity of the relations is strong for larger values of $\widehat{\gamma}_{\mathrm{ML}}$. This may be shown from the formula in section 2.3.3. Near zero the relations become messy, but this is not relevant for tests with normal 5% size. For this reason, tests that simply use the largest root of the score function perform equally well as the formal maximum likelihood ratio tests.

2.4 Unit root tests in the AR(1) model with serial correlation

2.4.1 General marginal likelihood tests

Until now we have explored tests for unit roots under the assumption that the error term v_t in (2.2) is serially uncorrelated. In practice in many cases this assumption is not valid. It is well known that misspecification of v_t has serious consequences for unit root tests. Notoriously difficult is the ARMA(1,1) (autoregressive moving average) model, where $v_t = \varepsilon_t - \theta\varepsilon_{t-1}$. Applying the AR(1) marginal likelihood based tests when data is generated by an ARMA(1,1)

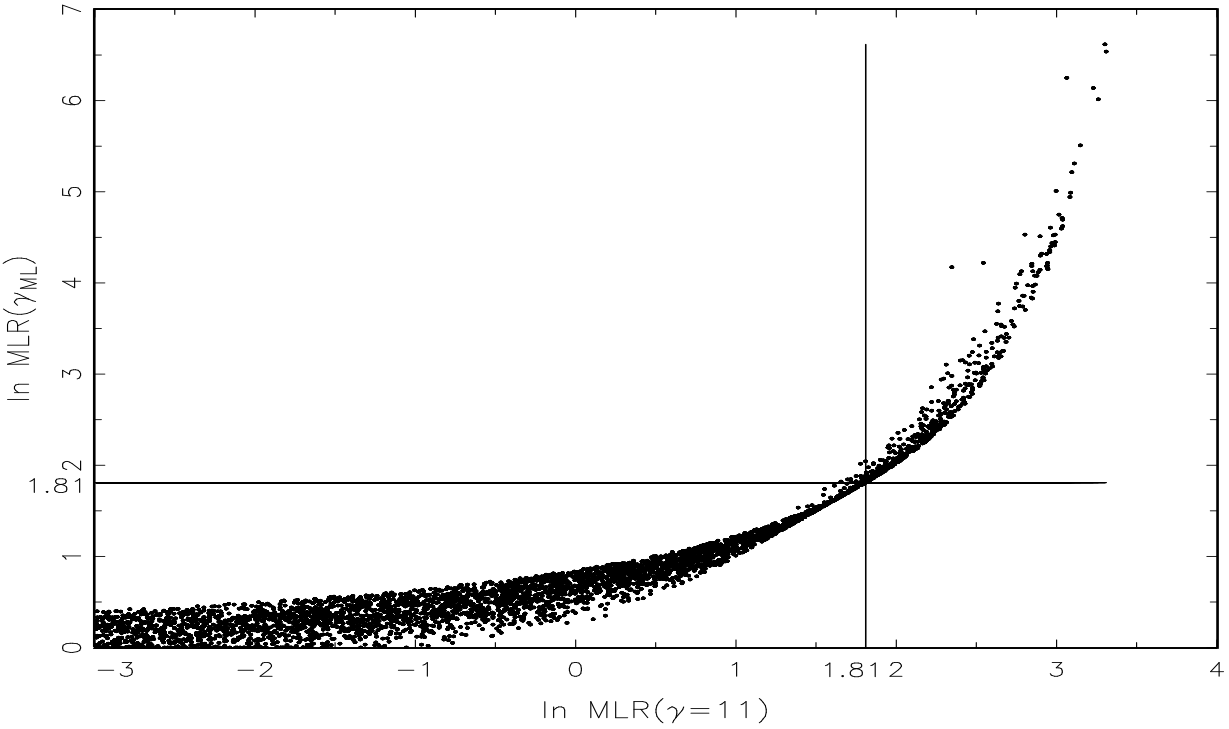

Figure 2.4: MLR($\gamma = 11$) versus MLR($\widehat{\gamma}_{ML}$).

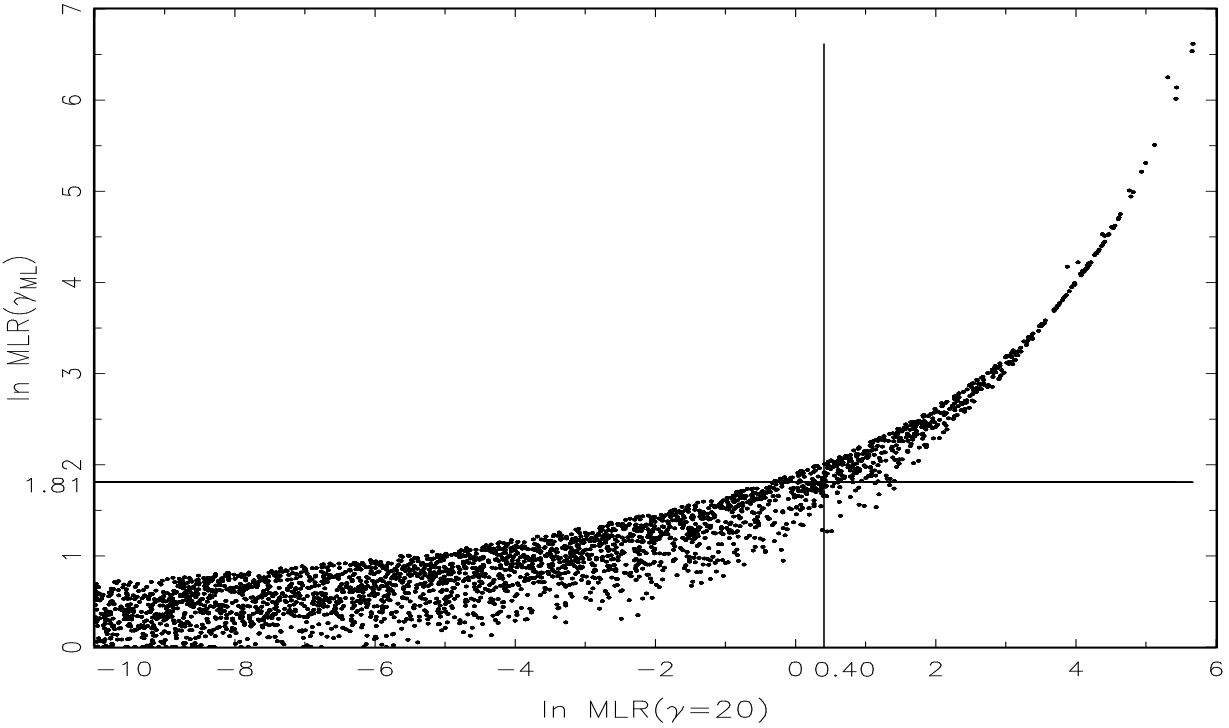

Figure 2.5: MLR($\gamma = 20$) versus MLR($\widehat{\gamma}_{ML}$).

process, results in very serious size distortions. This is already the case in the model without explanatory variables. For instance, the MLR test T_1 when $\theta = 0.5$ and $n = 100$, has a size of 0.62 when the 5 percent critical value (1.799 from table 2.2) is used.

There are two ways to cope with serial correlation in v_t within the context of marginal likelihood. The first is to apply standard likelihood methods for model selection and estimation. The marginal likelihood is well defined for relevant models. If in (2.2) v_t is a stationary ARMA(p,q) process, the marginal likelihood is also defined when $\rho = 1$, as $\Delta y_t = \Delta x_t\beta + \Delta u_t$, and $\Delta u_t = (1-L)/(1-\rho L)v_t$ is stationary for $-1 < \rho \leq 1$. Computations can be done efficiently by the exact initial Kalman filter, see Koopman (1997) and the diffuse Kalman filter, see De Jong (1988) and De Jong (1991a). Numerical problems in the computations of the Kalman filter for ρ near 1 can be avoided by formulating the model directly in first differences.

Model selection can be done by the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC) or other criteria. Conditional on the chosen model a MLR test may be applied on the unit root hypothesis. In the next subsection we will give the outcomes for the ARMA(1,1) case conditional upon the correct model choice. This provides a benchmark for the second approach.

The second approach is based on the asymptotic distribution of the MLR (2.19) under serial correlation. The asymptotically relevant parameters are the long term variance ω^2, and the unconditional variance of v_t, $\gamma_v(0)$ as defined in section 2.3.3. Asymptotically, apart from γ the MLR only depends on one parameter, $\kappa = \omega^2/\gamma_v(0)$. From equations (2.19) and (2.20) it follows that

$$\ln h^i(\gamma) + \kappa^{-1}\left(m_i \ln \frac{\mathrm{RSS}_i(\gamma)}{\mathrm{RSS}_i(0)} + \gamma\right) - \gamma \Rightarrow \ln h^i_\infty(\gamma) + \left(\gamma^2 A_\infty + \gamma\left(D^2_\infty - 1\right) - R^i_\infty\right), \quad (2.22)$$

where $m_i = n - k_i$, $i = \mu, \tau$, $k_\mu = 1$, and $k_\tau = 2$. The right-hand side has the distribution that is known from the AR(1) model where the v_t are i.i.d. Gaussian with variance σ^2. Consequently the adjusted residual sum of squares may be used in combination with the critical values computed for the marginal likelihood tests without correlation.

So, asymptotically, procedures based on estimation of κ will be robust. Estimation of κ is consistently possible, though outside the likelihood context, by ordinary least squares on

$$\Delta y_t = \alpha + \delta_0 y_{t-1} + x_t'\beta + \sum_{i=1}^{l} \delta_i \Delta y_{t-i} + \eta_t, \quad (2.23)$$

where l is the number of lags, following the approach of Elliott, Rothenberg, and Stock (1996). It follows that

$$\widehat{\kappa} = \frac{\hat{\omega}^2}{\hat{\gamma}_v(0)} = \sum_{i=l+2}^{n} \hat{\eta}_t^2 \left(1 - \sum_{i=1}^{l} \hat{\delta}_i\right)^{-2} \Big/ \sum_{i=l+2}^{n} \left(\Delta y_t - \hat{\alpha} - \widehat{\delta}_0 y_{t-1}\right)^2, \quad (2.24)$$

where $\hat{\omega}^2$ is an estimate of the spectral density at frequency 0, and $\hat{\gamma}_v(0)$ is an estimate of the unconditional variance of v_t. Lag length selection (l) can be done by information criteria. Ng and Perron (2001) show that AIC and BIC tend to select a lag that is too small, and they propose a modified information criterion. Alternatively nonparametric methods using for instance a Bartlett window can be applied to obtain an estimate of $\hat{\omega}^2$. Conditional on the estimate of κ the maximum marginal likelihood estimator of γ can be computed from the left-hand side of (2.22).

2.4.2 Simulation results for the ARMA(1,1) model

In this section we explore test results for the ARMA(1,1) model, $y_t = \mu + x_t'\beta + u_t$, $u_{t+1} = \rho u_t + \varepsilon_t - \theta\varepsilon_{t-1}$. In state-space format the model in first differences is provided by

$$\Delta y_t = \begin{pmatrix} 1 & 0 & 0 \end{pmatrix} \alpha_t + \Delta x_t \beta, \tag{2.25}$$

$$\alpha_{t+1} = \begin{pmatrix} \rho & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} \alpha_t + \begin{pmatrix} 1 \\ -(1+\theta) \\ \theta \end{pmatrix} \varepsilon_t, \tag{2.26}$$

for $t = 2, \ldots, n$, and $\alpha_1 \sim N(0, \sigma^2 P_1)$ where

$$P_1 = \begin{pmatrix} 2(\theta^2 + \theta + 1 - \rho\theta)/(1+\rho) & -(1+\theta)^2 + \theta\rho & -\theta \\ -(1+\theta)^2 + \theta\rho & 1 + 2\theta + 2\theta^2 & -\theta(1+\theta) \\ -\theta & -\theta(1+\theta) & \theta^2 \end{pmatrix}. \tag{2.27}$$

The marginal likelihood is computed by the Kalman filter and depends on 2 parameters, ρ and θ. The maximum marginal likelihood estimator under the null and alternative hypotheses can be substituted in for θ.

Unit root tests are applied, based on the marginal likelihood ratio ($T_1^{\hat{\theta}_{\text{ML}}}$), and on the difference between $\widehat{\rho}_{\text{ML}}$ and 1 ($T_2^{\hat{\theta}_{\text{ML}}}$), where $\widehat{\rho}_{\text{ML}}$ and $\hat{\theta}_{\text{ML}}$ are maximum marginal likelihood estimators from model (2.25)–(2.27). In table 2.6 the results from a small simulation study are presented for $\theta = 0.5$ with $n = 100$ and 1,000 observations, based on 10,000 replications. As critical values the 5 percent values from table 2.2 and 2.3 are used.

For $n = 100$ it can be concluded that the test $T_2^{\hat{\theta}_{\text{ML}}}$ suffers from size distortion. For $T_1^{\hat{\theta}_{\text{ML}}}$ the size distortion is small. The power of the tests $T_1^{\hat{\theta}_{\text{ML}}}$ and $T_2^{\hat{\theta}_{\text{ML}}}$ drops down for $\gamma = 20$ and $n = 100$ from 0.97 to 0.59 in the constant case, and from 0.78 to 0.32 in the constant and linear trend case, compared to the model without serial correlation. For $n = 1{,}000$ the power functions for $T_1^{\hat{\theta}_{\text{ML}}}$ and $T_2^{\hat{\theta}_{\text{ML}}}$ almost coincide with the asymptotic power envelopes for the AR(1) model.

The other test results of T_1 and T_2 in table 2.6 are based on the adjusted marginal likelihood ratio, as provided in (2.22). The tests based on knowledge of the true $\kappa = (1-\theta)^2/(1+\theta^2)$ are indicated by the superscript κ. The superscript $\hat{\kappa}$ means that κ is estimated in each replication

Table 2.6: Power functions for unit root test in the ARMA(1,1) model.

$\gamma = n(1-\rho)$		0	5	10	15	20
constant	n	power function				
T_1^{κ}	100	0.038	0.146	0.417	0.726	0.910
T_2^{κ}		0.040	0.153	0.438	0.758	0.935
$T_1^{\widehat{\kappa}}$		0.125	0.367	0.671	0.856	0.913
$T_2^{\widehat{\kappa}}$		0.128	0.372	0.686	0.870	0.924
$T_1^{\hat{\theta}_{\mathrm{ML}}}$		0.046	0.166	0.354	0.521	0.587
$T_2^{\hat{\theta}_{\mathrm{ML}}}$		0.079	0.257	0.522	0.716	0.807
$Q_T(10)$		0.164	0.467	0.766	0.903	0.958
T_1^{κ}	1000	0.051	0.202	0.515	0.827	0.964
T_2^{κ}		0.055	0.210	0.526	0.829	0.970
$T_1^{\widehat{\kappa}}$		0.077	0.290	0.653	0.893	0.987
$T_2^{\widehat{\kappa}}$		0.076	0.290	0.654	0.893	0.989
$T_1^{\hat{\theta}_{\mathrm{ML}}}$		0.057	0.224	0.543	0.824	0.963
$T_2^{\hat{\theta}_{\mathrm{ML}}}$		0.058	0.229	0.550	0.832	0.970
$Q_T(10)$		0.087	0.319	0.680	0.911	0.985
constant and trend						
T_1^{κ}	100	0.021	0.047	0.132	0.316	0.551
T_2^{κ}		0.029	0.061	0.166	0.372	0.624
$T_1^{\widehat{\kappa}}$		0.176	0.276	0.515	0.752	0.897
$T_2^{\widehat{\kappa}}$		0.183	0.289	0.537	0.774	0.915
$T_1^{\hat{\theta}_{\mathrm{ML}}}$		0.045	0.080	0.168	0.271	0.341
$T_2^{\hat{\theta}_{\mathrm{ML}}}$		0.095	0.174	0.336	0.509	0.633
$Q_T(10)$		0.234	0.381	0.615	0.795	0.899
T_1^{κ}	1000	0.046	0.092	0.233	0.481	0.742
T_2^{κ}		0.048	0.095	0.237	0.487	0.751
$T_1^{\widehat{\kappa}}$		0.041	0.081	0.198	0.409	0.652
$T_2^{\widehat{\kappa}}$		0.042	0.082	0.202	0.411	0.659
$T_1^{\hat{\theta}_{\mathrm{ML}}}$		0.049	0.099	0.250	0.483	0.732
$T_2^{\hat{\theta}_{\mathrm{ML}}}$		0.053	0.107	0.264	0.505	0.750
$Q_T(10)$		0.107	0.198	0.417	0.690	0.871

T_i^{κ} is the test based on the adjusted marginal likelihood ratio (2.22), using the true value of κ.
$T_i^{\widehat{\kappa}}$ is the test based on the adjusted marginal likelihood ratio (2.22), where κ is estimated by (2.24).
$T_i^{\hat{\theta}_{\mathrm{ML}}}$ is the test, where γ and θ are estimated from (2.25)–(2.27), $i = 1, 2$.
The data generating process is provided by (2.1)–(2.3), where $v_t = \varepsilon_t - \theta\varepsilon_{t-1}$, $\theta = 0.5$, and $\sigma_\varepsilon^2 = 1$.

by (2.24). Like in the $Q_T(10)$ test the lag length l is determined by BIC with a maximum lag of 4. For the ARMA(1,1) model with κ known the results are satisfactory for both $n = 100$ and $n = 1{,}000$. For the case where κ is estimated, there is considerable size distortion, diminishing from 0.13 for $n = 100$ to 0.08 for $n = 1{,}000$ for the model with constant, and diminishing from 0.18 for $n = 100$ to 0.04 for $n = 1{,}000$ for the model with constant and trend. These results are better than the results for the $Q_T(10)$ test, but worse than the tests T_1 and T_2 (to be expected as the model is assumed known), specifically T_1. From table 4 in Elliott (1999) it appears that other tests like the ADF, DF-GLSu^d, P_T, and the CWSSE tests, have size distortions in the same range: from 0.18 to 0.23 for the constant case, and 0.25 to 0.33 for the constant and linear trend case, for $n = 100$.

The overall picture is satisfactory compared to the literature. And there are many possibilities for improvement, specifically in the estimation procedure for κ.

2.5 Conclusion

In the general linear model with a disturbance structure that possibly contains a unit root, the profile likelihood degenerates in the unit root. The reason is that the initial condition is unknown for $\rho = 1$. This problem disappears if the marginal likelihood is used in combination with a specification that has an unknown equilibrium level together with a covariance stationary specification of the disturbance term for $\rho < 1$, and an unknown initial condition for $\rho = 1$. As the marginal likelihood is the likelihood of the maximal invariant, the general optimality properties of maximum likelihood inference apply. Maximum marginal likelihood estimation is possible, e.g., for complex correlation structures and/or nonnormal disturbances. Standard AIC or BIC criteria may be used for model selection.

The asymptotic behaviour of marginal likelihood ratio tests is evaluated in the local-to-unity framework, $\gamma = n(1-\rho)$. For the AR(1) model with constant (and trend), the MLR tests appear to perform close to their theoretical upper bound, asymptotically as well as in finite samples. The asymptotic critical values can even be used for samples as small as $n = 25$.

If the disturbance term has an ARMA(1,1) structure, the maximum marginal likelihood ratio test appears to outperform other tests, specifically with respect to size in small samples. For general correlation structures, there is an alternative for formulating the marginal likelihood of fully specified models. This is based on the well known asymptotic results for non-stationary processes. We derive an adjustment of the standard MLR test, which depends only on the ratio of the long term variance and the unconditional variance. The adjusted MLR has an asymptotic distribution that only depends on γ.

We have carried out a simulation study for the ARMA(1,1) model. The power functions for sample size $n = 1{,}000$ show that the asymptotic adjustment works well, for $n = 100$ the results outperform other tests.

Chapter 3

Marginal likelihood, Jeffreys' rule and unit root tests

Abstract

In inference on the covariance matrix of the general linear model, location and scale parameters are nuisance. Classical marginal likelihood is based on a transformation of the data that removes these parameters. Bayesian marginal likelihood is obtained by integrating out these parameters. First it is shown that both likelihoods are proportional when the independence Jeffreys' prior is used, which differs from Jeffreys' rule.

The major difference between Bayesian and classical inference exists in the context of testing. However, it is shown that in case of a monotone marginal likelihood ratio depending on only one parameter, the marginal likelihood ratio test and the Bayesian posterior odds test use the data exactly in the same way. Using proper priors, the only difference for one-sided hypothesis tests is that the size of the posterior odds test is determined by prior considerations. As the marginal likelihood ratio test is uniformly most powerful invariant, the same holds for the posterior odds test.

These two results enable us to show a strong analogy between classical and Bayesian unit root testing in the linear model with first order autoregressive disturbances. In the previous chapter it was shown that the marginal likelihood ratio is approximately monotone. As a consequence the posterior odds test has the same form as from the marginal likelihood ratio test, when the independence Jeffreys' prior is used.

The implied size of the posterior odds test strongly depends on the prior for the autoregressive parameter ρ. It is shown that priors in terms of ρ imply a relation between the test size and the sample size, opposed to priors formulated in the local-to-unity format $\gamma = n(1-\rho)$.

3.1 Introduction

This chapter compares classical and Bayesian inference on parameters in the covariance matrix in the general linear model. The location and scale parameters are considered as nuisance parameters.

Marginal likelihood is a term used in both classical and Bayesian statistics. The classical concept is based on a transformation of the data to remove nuisance parameters. In Bayesian statistics, nuisance parameters are integrated out. In the general linear model both procedures are feasible with respect to location and scale parameters. Unlike the prior following from Jeffreys' rule, the independence Jeffreys' prior appears to establish proportionality between both likelihoods. We argue that there is a strong case to use this prior (or even classical marginal likelihood directly) in a Bayesian analysis. Consequently classical tests and Bayes factors can be based on the same marginal likelihood ratio.

There are major differences between classical and Bayesian inference in the context of hypothesis testing. However, in case of a marginal likelihood depending on only one parameter and a monotone marginal likelihood ratio, a further correspondence between classical and Bayesian analysis exists. We consider one-sided hypothesis tests and proper priors for the parameter of interest. In that case marginal likelihood ratio tests and posterior odds tests use the data in the same way. The only difference concerns the size of the test. In the classical framework it is a prechosen value. In the Bayesian analysis it follows from prior considerations: the prior for the parameter concerning the covariance structure, prior odds, and the loss function. Different priors result in different sizes. In case of a monotone marginal likelihood ratio the marginal likelihood ratio test is uniformly most powerful invariant (UMPI), and it will be shown the same holds for the posterior odds test.

Both results, the proportionality between the Bayesian and classical marginal likelihood and a similar use of the data in the testing procedures, provide the basis to show a strong analogy between classical and Bayesian unit root testing in the linear model with first order autoregressive disturbances.

There is a large amount of literature on Bayesian unit root testing and differences with the classical approach. The many options in model specification, prior distribution and the treatment of the initial condition make this literature rather complex. The special issues of the *Journal of Applied Econometrics* (1991) and the *Journal of Econometrics* (1995) were devoted to the comparison of the two approaches. The power of classical unit root tests in small samples was questioned by Sims (1988) from a Bayesian point of view. Phillips (1991) claims that the difference in results between classical and Bayesian inference is a result of the use of a flat prior for the autoregressive parameter. Sims and Uhlig (1991) have designed an experiment to compare Bayesian and classical inference, although in a model without a constant, different from (2.10)–(2.12). As shown by Lubrano (1995) the choice of the model and the treatment of the initial condition is essential.

In this chapter we restrict ourselves to the model with first order autoregressive disturbances, $y_t = \rho y_{t-1} + (1-\rho)\mu + \varepsilon_{t-1}$, as provided in (2.10)–(2.12) in the unobserved component format. In the previous chapter it was shown that the classical marginal likelihood depends on only one parameter, is nonzero and finite in $\rho = 1$ and continuous for $\rho \uparrow 1$, and that tests based on the marginal likelihood are almost UMPI. Bayesian marginal likelihood has been used for the same

problem, see for example Zellner (1971) and Lubrano (1995). However, due to the fact that the data are informative on $\mu(1-\rho)$, Jeffreys' rule leads to a prior $\pi(\mu,\sigma^2|\rho)$ containing a factor $(1-\rho)$ which leads, in combination with a proper prior for ρ, to a posterior that is zero for $\rho=1$. The use of the independence Jeffreys' prior removes this problem. An extensive survey of singularities at $\rho=1$ for different choices of priors, including the independence Jeffreys' prior, called "flat" prior, model specifications, and initial conditions is provided by Bauwens, Lubrano, and Richard (1999, ch. 6).

In the AR(1) model with constant the marginal likelihood ratio is approximately monotone in $\widehat{\rho}_{\mathrm{ML}}$, see section 2.3.5. It will be shown that for that reason the posterior odds test has the same power as the marginal likelihood ratio test. The implied size depends on the specification of the prior for ρ and on the sample size n. The marginal likelihood ratio has a limiting distribution under the null hypothesis in the local-to-unity format $\gamma=n(1-\rho)$, see section 2.3.3. As a consequence, when priors are formulated in terms of γ instead of ρ, the implied size of the test is independent of the sample size.

The setup of the chapter is as follows. Section 3.2 considers the relation between Bayesian and classical marginal likelihood in the general linear model. In section 3.3 the correspondence between posterior odds and marginal likelihood ratio tests is treated. Section 3.4 compares classical and Bayesian inference on unit root tests. The size of the posterior odds test is studied for different priors and varying sample size. Section 3.5 concludes.

3.2 Jeffreys' rule and marginal likelihood

In Bayesian inference the term marginal likelihood directly follows from the definitions of probability calculus. In the general linear model where β and σ^2 are nuisance parameters, the Bayesian marginal likelihood is provided by

$$f(y|\theta)=\int\int f(y|\theta,\beta,\sigma^2)\pi(\beta,\sigma^2|\theta)\mathrm{d}\beta\mathrm{d}\sigma^2, \tag{3.1}$$

where $\pi(\beta,\sigma^2|\theta)$ is a prior. The posterior of θ follows from $f(\theta|y)\propto f(y|\theta)\pi(\theta)$: marginal posterior is proportional to marginal prior times marginal likelihood.

If Bayesians would claim the term "marginal likelihood" to prevent confusion, they would have a strong case. However, they should always add to marginal likelihood "for a given prior $\pi(\beta,\sigma^2|\theta)$". The use of the term marginal likelihood without reference to a prior should be reserved for the case $\pi(\beta,\sigma^2|\theta)$ is noninformative. Unfortunately this is only possible for degenerate priors in which case $f(y|\theta)$ is not a proper likelihood. Moreover the definition of noninformative is anything but settled.

In section 2.2.1 it is shown that y^* is a maximal invariant for the transformations (2.9) and it is argued that the classical marginal likelihood $f(y^*|\theta)$ contains all information on θ in

absence of knowledge of β and σ^2. The operational definition to incorporate this is

$$f(\theta|y) = f(\theta|y^*). \tag{3.2}$$

Consequently,

$$f(\theta|y) = f(\theta|y^*) \propto f(y^*|\theta)\pi(\theta). \tag{3.3}$$

It provides a direct way to avoid improper priors for the nuisance parameters β and σ^2 in a Bayesian analysis, and no problems with a degenerate likelihood arise. One simply uses the likelihood of the transformed data $y^* = A'y/\sqrt{y'AA'y}$ as provided in (2.8).

Proportionality between the Bayesian and classical marginal likelihood is established by the use of the noninformative prior

$$\pi(\beta, \sigma^2|\theta) = \pi(\beta, \sigma^2) \propto \sigma^{-2}. \tag{3.4}$$

We refer to this prior as the independence Jeffreys' prior, because it implies that

$$\pi(\beta, \sigma^2, \theta) \propto \pi(\beta)\pi(\sigma^2)\pi(\theta) \propto \sigma^{-2}\pi(\theta). \tag{3.5}$$

The proof that this prior makes Bayesian and classical marginal likelihood proportional is simple algebra. Define $B^{*\prime}y$ as the complement of y^*, such that

$$f(y|\theta, \sigma^2, \beta) = f(y^*|\theta)f(B^{*\prime}y|\theta, \sigma^2, \beta), \tag{3.6}$$

then

$$f(B^{*\prime}y|\theta, \sigma^2, \beta) \propto \left|X'\Omega^{-1}X\right|^{1/2} |\Omega|^{1/2} (y'\Omega^{-1}M_X^{\Omega}y)^{m/2} f(y|\theta, \sigma^2, \beta), \tag{3.7}$$

and

$$\int\int f(B^{*\prime}y|\theta, \sigma^2, \beta)\pi(\beta, \sigma^2|\theta)\mathrm{d}\beta\mathrm{d}\sigma^2 \propto 1, \tag{3.8}$$

for the noninformative prior (3.4). From (3.1), (3.6), and (3.8) it follows that

$$f(y|\theta) \propto f(y^*|\theta). \tag{3.9}$$

The remaining question is whether $\pi(\beta, \sigma^2|\theta) \propto \sigma^2$ is the only one that leads to $f(\theta|y) = f(\theta|y^*)$. Appendix A.3 shows for the general linear model that this is the case, at least for the class of conjugate priors that contains all suggestions ever made for noninformative priors.

The prior (3.4) differs from the prior following from Jeffreys' rule, that says that the prior is proportional to the square root of the determinant of the Fisher information matrix associated with the likelihood function of the model. This rule is usually applied in univariate cases and its application in multivariate cases yields unwanted results. The strictly use of Jeffreys' rule in the simple linear regression model without covariance structure would lead to a prior $\pi(\beta, \sigma^2) \propto$

$\sigma^{-(k+2)}$. As this has implausible consequences, Jeffreys assumed a priori independence between $\beta|\sigma^2$ and σ^2 to obtain $\pi(\beta, \sigma^2) \propto \sigma^{-2}$, according to Bernardo and Smith (1994, p. 361) an *ad hoc* recommendation (italics from Bernardo and Smith). The application of Jeffreys' rule separately to σ^2 and (β,θ), may lead to further problems, as will become clear in the discussion of the AR(1) model. Problems that are solved, when (3.4) is used.

Reference analysis, see Berger and Bernardo (1992), does not lead to unambiguous results. Fernández and Steel (1999) show that for inference on scale and location parameters the reference prior equals $\pi(\beta, \sigma^2) \propto \sigma^{-2}$. However, when θ is included, reference analysis becomes complicated and ambiguous. We discuss this later for the AR(1) model.

A point of attention is that equation (3.4) refers to priors that are meant to be noninformative with respect to inference on θ. It should not automatically be used for other inference, as the purpose of inference matters for the choice of reference priors, see Stone and Dawid (1972).

The conclusion is that classical marginal likelihood can be used for inference on θ in an "objective" Bayesian framework, and that there is a plausible Bayesian derivation that leads to equivalent results. The classical formulation (2.8), using the proportionality constant $|X'X|^{-1/2}$ and a well defined probability function, may have advantages, even for Bayesians. For inference on θ it does not matter, for other inference like model comparison it is more informative and avoids errors that are easily made using improper distributions.

3.3 Bayes factors and classical tests

In this section we compare classical and Bayesian hypothesis testing in the general linear model, where the covariance parameter θ is scalar. The parameters β and σ^2 are nuisance. We consider the one-sided test $H_0 : \theta = \theta_0$, against an alternative $H_1 : \theta > \theta_0$ (or $\theta < \theta_0$).

There are important differences between the Bayesian and classical hypothesis testing approaches. In contrast to the classical approach, Bayesian hypothesis testing procedures requires a prior distribution on θ, and a loss function. In this one-sided test context, we have to use a mixed prior probability for θ. We restrict ourselves to the use of proper priors for θ.

Despite the major differences between the two approaches we will show in this section that in case of a likelihood depending on only one parameter, and a monotone likelihood ratio, the Bayesian and classical approach have much in common.

3.3.1 Classical hypothesis testing

Classical hypothesis testing can be based on the likelihood ratio. The likelihood ratio contains the nuisance parameters β and σ^2. In section 2.2.1 it is shown that the marginal likelihood $L_{M_{\beta,\sigma}}(\theta)$ can be treated as the likelihood function of θ. As the marginal likelihood contains only one parameter and the marginal likelihood ratio is monotone in some statistic S, the test that rejects H_0 if $S(y) > \kappa_0^*$, is uniformly most powerful invariant (UMPI), see Lehmann (1986). In

general this is the case if S is a sufficient statistic. The marginal likelihood ratio test evaluated in the maximum likelihood estimator of θ is also UMPI, because $\text{MLR}(\widehat{\theta}_{\text{ML}})$ is a function of S. The MLR test has the format: reject H_0 if

$$\text{MLR}(\widehat{\theta}_{\text{ML}}) = \frac{f(y^*|\widehat{\theta}_{\text{ML}})}{f(y^*|\theta_0)} > \kappa_0, \tag{3.10}$$

where κ_0 is chosen such that

$$P_{y^*|H_0}(\text{MLR}(\widehat{\theta}_{\text{ML}}) > \kappa_0) = \alpha, \tag{3.11}$$

and α is the predetermined size of the test.

Alternatively, the p-value can be calculated as

$$p = P_{y^*|H_0}(\text{MLR}_{y^*}(\widehat{\theta}_{\text{ML}}) > \text{MLR}_{\underline{y}^*}(\widehat{\theta}_{\text{ML}})), \tag{3.12}$$

where $\text{MLR}_{\underline{y}^*}$ is the observed marginal likelihood ratio. The null hypothesis is rejected if $p < \alpha$.

3.3.2 Bayesian hypothesis testing

Bayesian tests are based on posterior odds. The posterior odds ratio can be expressed as prior odds times Bayes factor,

$$\frac{f(H_1|y)}{f(H_0|y)} = \frac{\pi(H_1)}{\pi(H_0)} \frac{f(y|H_1)}{f(y|H_0)}, \tag{3.13}$$

A full Bayesian motivation of the choice between $\theta = \theta_0$ and the alternative thus requires the specification of a prior $\pi(\theta|H_1)$, prior odds and a loss function such that the Bayes Factor may be used to decide whether the decision $\theta = \theta_0$ is better than the alternative.

The link to the classical choice between hypotheses is provided by the observation that a Bayesian decision rule has the format "Choose H_1 (reject H_0) if $\text{BF} > \kappa$". Given a loss function $L(i, j)$ when model i is chosen while model j is true, one must choose H_1 if

$$\text{BF} = \frac{f(y|H_1)}{f(y|H_0)} > \frac{\pi(H_0)}{\pi(H_1)} \frac{L(1,0)}{L\,(0,1)} = \kappa. \tag{3.14}$$

The Bayes factor in the one-sided test is given by

$$\text{BF} = \frac{\int_{H_1} f(y|\theta)\pi(\theta|H_1)\text{d}\theta}{\int_{H_0} f(y|\theta)\pi(\theta|H_0)\text{d}\theta} = \int_{H_1} \frac{f(y|\theta)}{f(y|\theta_0)} \pi(\theta|H_1)\text{d}\theta, \tag{3.15}$$

a weighted average of likelihood ratio's. Formally, the Bayes factor is not defined when an improper prior $\pi(\beta, \sigma^2|\theta)$ is used to derive the marginal likelihood $f(y|\theta)$: the marginal likelihood is known up to a proportionality constant. This problem can be circumvented by using the

classical marginal likelihood directly to obtain the Bayes factor,

$$\mathrm{BF} = \int_{H_1} \frac{f(y^*|\theta)}{f(y^*|\theta_0)} \pi(\theta|H_1) \mathrm{d}\theta, \tag{3.16}$$

as $f(y^*|\theta)$ is a well defined density function and proportional to $f(y|\theta)$ for the prior (3.4).

Although it is no common practice in a Bayesian analysis it is possible to compute the probability of the "type I" error. The outcome of the Bayesian decision rule (3.14) corresponds to a classical test with size α,

$$P_{y^*|H_0} \left(\int_{H_1} \frac{f(y^*|\theta)}{f(y^*|\theta_0)} \pi(\theta|H_1) \mathrm{d}\theta > \kappa \right) = \alpha. \tag{3.17}$$

The α-level does not depend on the data, and can be reconstructed from κ and $\pi(\theta|H_1)$ analytically or by simulation.

Dual to the classical hypothesis testing procedure it is possible to compute a "p-value". Define $\mathrm{BF}_{\underline{y}^*}$ as the observed Bayes factor and BF_{y^*} as a possible outcome when y^* is generated under the null hypothesis. The Bayesian p-value is simply

$$p = P_{y^*|H_0}(\mathrm{BF}_{y^*} > \mathrm{BF}_{\underline{y}^*}). \tag{3.18}$$

Under the null it has a uniform $U(0;1)$ distribution. Although both BF_{y^*} and $\mathrm{BF}_{\underline{y}^*}$ depend on the prior $\pi(\theta|H_1)$, the p-value does not. It only depends on the data. This result follows from the monotone marginal likelihood ratio. The marginal likelihood ratio $\mathrm{MLR}(\theta)$ can be expressed as $f(\theta, S)$, where f is monotone in $S = S(y^*)$, and therefore $g(S) = \int_{H_1} f(\theta, S)\pi(\theta|H_1)\mathrm{d}\theta$ is monotone in S as well. As as consequence the p-value (3.18) can be expressed as

$$\begin{aligned} p &= P_{y^*|H_0}(g(S(y^*)) > g(S(\underline{y}^*))) = P_{y^*|H_0}(S(y^*) > S(\underline{y}^*)) \\ &= P_{y^*|H_0}(f(\theta, S(y^*)) > f(\theta, S(\underline{y}^*))) = P_{y^*|H_0}(\mathrm{MLR}_{y^*}(\widehat{\theta}_{\mathrm{ML}}) > \mathrm{MLR}_{\underline{y}^*}(\widehat{\theta}_{\mathrm{ML}})). \end{aligned} \tag{3.19}$$

The Bayesian p-value decision rule can be formulated as "Choose H_1 (reject H_0) if $p < \alpha$". In a classical analysis the size α of the test is predetermined. In a Bayesian analysis it implicitly follows from the prior $\pi(\theta|H_1)$ and κ, see equation (3.17). In case of a monotone likelihood ratio it is possible to derive the implied value of κ, given α and $\pi(\theta|H_1)$. The relation between κ and α strongly depends on $\pi(\theta|H_1)$. From the monotonicity of g it follows that

$$P_{y^*|H_0}(g(S) > g(\psi_\alpha)) = P_{y^*|H_0}(S > \psi_\alpha), \tag{3.20}$$

where ψ_α is the α-critical value of S under H_0. Consequently,

$$\kappa = \int f(\theta, \psi_\alpha)\pi(\theta|H_1)\mathrm{d}\theta, \tag{3.21}$$

Table 3.1: The relation between the κ and α analysis.

	statistic	threshold
κ analysis	$\mathrm{BF}(y^*, \pi(\theta\|H_1))$	κ
α analysis	$P_{y^*\|H_0}(\mathrm{BF}_{y^*} > \mathrm{BF}_{\underline{y}^*})$	$\alpha(\kappa, \pi(\theta\|H_1))$

the integral over the α-quantiles of the marginal likelihood ratio function. In the next section this representation will prove to be very useful.

The result that $P_{y^*|H_0}(\mathrm{BF}_{y^*} > \mathrm{BF}_{\underline{y}^*}) = P_{y^*|H_0}(\mathrm{MLR}_{y^*}(\widehat{\theta}_{\mathrm{ML}}) > \mathrm{MLR}_{\underline{y}^*}(\widehat{\theta}_{\mathrm{ML}}))$ is related to Andrews (1994), who showed – under more general assumptions – that for certain priors the Bayesian posterior odds tests is equivalent in large samples to classical likelihood ratio tests with a size determined by prior considerations.

A final remark is that as the marginal likelihood ratio test is UMPI, by (3.19) the same holds for the Bayesian posterior odds test.

3.3.3 The use of the p-value

In the previous subsection two different Bayesian decision rules were presented: the standard approach using Bayes factors, and the "p-value" approach. Table 3.1 shows both representations of the posterior odds tests, the κ representation (3.14) and the α representation (3.18).

Bayesians normally do not compute p-values. In case of a monotone likelihood ratio and a likelihood containing only one parameter, we think there are good reasons to do it. Choosing the p-value as the test statistic to communicate, has a number of advantages. It facilitates the discussion with frequentists, and even for a Bayesian it might be interesting to derive the probability of the "type I" error. Another advantage is that the evidence from the data and prior considerations are separated, which is useful for sensitivity analysis. The p-value provides all relevant information from the data. The discussion on the appropriate priors and loss function to determine whether the statistic is sufficiently informative to decide against the null is a separate and subjective matter, where readers can make different choices.

On the relevant value of α there is substantial discussion in the literature. It is well known that for $\kappa = 1$ (a default choice for most Bayesians) sizes of at least about 0.25 are needed, see Berger (2003) for references. This choice of κ however is in no way compulsory and may differ from situation to situation, depending on prior beliefs and loss functions. One may just as well argue that, in absence of a context, the default choice of $\alpha = 0.05$ has, notwithstanding many cases where it is inappropriate or interpreted badly, proven to be reasonable.

This setup also sheds some light on an old discussion. Berger (2003) gives an overview of the (dis)agreements between the arguments to compute p-values, advocated by Fisher because they are an index of the strength of evidence against the null, and the prechosen α-level advocated by Neyman because it satisfies the frequentist principle that in repeated use of the test the average error should not be greater than the average reported error. In our simple setting,

the Bayesian view is that there is not such a thing as a prechosen α-level, while the p-value is relevant in all cases.

The asymptotic correspondence between Bayesian posterior odds and classical tests of some size for a much wider class of models has been shown by Andrews (1994). The preceding remarks thus apply approximately in a wider setting. Noteworthy is Andrews' statement (p. 1208) that his results "do not apply to tests of a unit root". However, in the next section it is shown that the same results for unit root tests holds approximately, though not based on asymptotics.

3.4 Classical and Bayesian testing for a unit root

3.4.1 Priors for ρ

In the previous sections it is shown that 1) classical and Bayesian marginal likelihood are proportional when the independence Jeffreys' prior is used, and 2) that the marginal likelihood ratio and the posterior odds test use the data in a similar way in case of a monotone marginal likelihood ratio depending on only one parameter. Both results provide the basis to show a strong analogy between classical and Bayesian unit root testing in the linear model with first order autoregressive disturbances, as provided in the previous chapter in (2.10)–(2.12).

The classical marginal likelihood for this model is given by

$$L_{M_{\beta,\sigma}}(\rho) = f(y^*|\rho) = \frac{\frac{1}{2}\Gamma(\frac{n-1}{2})}{\pi^{(n-1)/2}} \left(\frac{n(1+\rho)}{n-(n-2)\rho}\right)^{1/2} \left(\frac{\text{RSS}_\mu(\rho)}{\text{RSS}_\mu(0)}\right)^{-(n-1)/2}, \qquad (3.22)$$

where $\text{RSS}_\mu(\rho)$ is given in (2.15). As follows from section 3.2 the Bayesian marginal likelihood $f(y|\rho)$ is proportional to the classical one, when the noninformative prior $\pi(\mu, \beta, \sigma^2|\rho) \propto \sigma^{-2}$ is used. The Bayesian marginal likelihood was derived by Lubrano (1995, equation 22). For $|\rho| < 1$ this result is equal to (2.18) though with a unspecified constant term. He shows that this expression has a finite limit for $\rho \to 1$. However, different from (2.18) the marginal likelihood in $\rho = 1$ is not defined as the initial condition $y_1 \sim N(\mu + x_1'\beta, \sigma^2/(1-\rho^2))$ is degenerate in $\rho = 1$.

As appears from the sequel in Bauwens, Lubrano, and Richard (1999, ch. 6) the prior $\pi(\mu, \beta, \sigma^2|\rho) \propto \sigma^{-2}$ is not undisputed. Following Zellner (1971) Jeffreys' rule for the multiparameter case applied separately to σ^2 and the other parameters, leads to $\pi(\mu, \beta, \sigma^2|\rho) \propto (1-\rho)\sigma^{-2}$. The term $(1-\rho)$ implies in combination with a proper prior for ρ, a posterior that is zero for $\rho = 1$. Bauwens, Lubrano, and Richard (1999) give a survey of singularities at $\rho = 1$ for different choices of priors, model specifications, and initial conditions. They recommend the nonlinear specification (2.10)–(2.12) in combination with the independence Jeffreys' prior.

That reference priors do not unambiguously lead to the same results, appears from Ghosh

and Heo (2003). They derived reference priors $\pi(\mu, \beta, \sigma^*, \rho)$ for inference on ρ in the model specified by (2.10)–(2.12), where $\sigma^* = \sigma^2(1-\rho^2)^{-1/n}$. It turns out to matter whether σ^* and (μ, β) are treated simultaneously or sequentially, in their notation π_{R2} and π_{R3}, respectively. The reference priors are provided by

$$\pi_{R2}(\mu, \beta, \sigma^*, \rho) \quad = \sigma^{*-3/2}(1-\rho^2)^{-1}\sqrt{n(1-\rho^2)+2\rho^2}, \tag{3.23}$$

$$\pi_{R3}(\mu, \beta, \sigma^*, \rho) \quad = \sigma^{*-1}(1-\rho^2)^{-1}\sqrt{n(1-\rho^2)+2\rho^2}. \tag{3.24}$$

It can be deduced that the conditional reference prior $\pi(\mu, \beta, \sigma^*|\rho)$ corresponds to the independence Jeffreys' prior (3.4) only for π_{R3}. If we would use the prior

$$\pi_{MR3}(\rho) \propto \sqrt{n(1-\rho^2)+2\rho^2} \tag{3.25}$$

in combination with the marginal likelihood (2.18) we would obtain their posterior (equation 15). One might expect that (3.25) equals the marginal reference prior $\pi_{R3}(\rho)$, but this is not the case as $\pi_{R3}(\rho) \propto (1-\rho^2)^{-1}\pi_{MR3}(\rho)$. Neither it is true that the reference prior derived from the marginal likelihood (2.18) equals (3.25). We did not derive this analytically but by simulation. The reason for the difference in priors is probably that Ghosh and Heo (2003) use a different transformation of the parameters. We did not pursue this further. Our approach is to concentrate on the marginal likelihood and to investigate the role of priors $\pi(\rho)$ separately, which will be done in the next subsection

3.4.2 Unit root tests

In this section we compare the power of marginal likelihood ratio tests with Bayesian posterior odds tests in the AR(1) model. The marginal likelihood ratio for this model is provided by

$$\mathrm{MLR}(\rho) = \frac{L_{M_{\beta,\sigma}}(\rho)}{L_{M_{\beta,\sigma}}(1)} = \left(\frac{1+\rho}{n(1-\rho)+2\rho}\right)^{1/2}\left(\frac{\mathrm{RSS}_\mu(\rho)}{\mathrm{RSS}_\mu(1)}\right)^{-(n-1)/2}, \tag{3.26}$$

where $\mathrm{RSS}_\mu(\rho)$ is provided in (2.15).

In section 2.3.3 it was demonstrated that even asymptotically the marginal likelihood ratio is a linear combination of more than one statistic, with weights that depend on γ. However, under the null hypothesis, $\mathrm{MLR}(\rho)$ is almost a monotone function of $\mathrm{MLR}(\widehat{\rho}_{\mathrm{ML}})$ for values of ρ not too close to $\rho = 1$, see section 2.3.5.

As the marginal likelihood ratio only depends on ρ and is approximately monotone in $\widehat{\rho}_{\mathrm{ML}}$, we might expect that the marginal likelihood ratio test and the posterior odds test have the same power function. In this subsection this is investigated for the AR(1) model. We expect that for the ARX(1) model similar results will apply, but we did not pursue this further.

Table 3.2: Power functions for the MLR test and Bayes Factors.

ρ	1.00	0.95	0.90	0.85	0.80
$P(\text{MLR} > \kappa_0 = 6.04)$	0.050	0.192	0.531	0.839	0.973
$P(\text{BF}_1 > \kappa_1 = 1.56)$	0.050	0.191	0.527	0.840	0.975
$P(\text{BF}_2 > \kappa_2 = 2.76)$	0.050	0.191	0.527	0.838	0.974
Power envelope	0.050	0.196	0.521	0.838	0.973

A Bayesian unit root test ($H_0 : \rho = 1$, and $H_1 : |\rho| < 1$) is based on the Bayes factor

$$\text{BF} = \frac{f(y^*|H_1)}{f(y^*|H_0)} = \int \frac{f(y^*|\rho)}{f(y^*|\rho = 1)} \pi(\rho|H_1) \mathrm{d}\rho. \tag{3.27}$$

If $\text{BF} > \kappa$ the alternative is chosen.

The classical test based on the marginal likelihood ratio has the format: reject H_0 if

$$\text{MLR}(\widehat{\rho}_{\text{ML}}) = \frac{f(y^*|\rho = \widehat{\rho}_{\text{ML}})}{f(y^*|\rho = 1)} > \kappa_0, \tag{3.28}$$

where κ_0 is chosen such that

$$P_{y^*|H_0}\left(\frac{f(y^*|\rho = \widehat{\rho}_{\text{ML}})}{f(y^*|\rho = 1)} > \kappa_0\right) = \alpha, \tag{3.29}$$

and α is the predetermined size of the test.

Here some results are provided for the equivalence between the Bayesian test in terms of κ and the classical marginal likelihood test in terms of α. As explained in section 3.3 this correspondence depends on $\pi(\rho|H_1)$. Table 3.2 compares power functions for three tests for $n = 100$ (10,000 replications): two Bayesian tests, based on a uniform and exponential prior and one classical marginal likelihood ratio test. The priors are

$$\pi_1(\rho) = U(0.5; 1), \tag{3.30}$$

$$\pi_2(\rho) = 6.9133 \times \exp(-20/3(1 - \rho)), \tag{3.31}$$

for $0.5 \leq \rho \leq 1$, with κ_i such that $P_{y^*|H_0}(\text{BF} > \kappa_i) = \alpha = 0.05$ for $i = 1, 2$. κ_0 follows from $P_{y^*|H_0}(\text{MLR}(\widehat{\rho}_{\text{ML}}) > \kappa_0) = \alpha = 0.05$. Note that κ_0 is a upperbound for values of κ obtained for different priors as it is based on the prior giving most weight to the alternative given the data: $\pi(\widehat{\rho}_{\text{ML}}) = 1$ for $\rho = \widehat{\rho}_{\text{ML}}$.

From table 3.2 it can be concluded that the power functions for the three tests are indistinguishable and very close to the power envelope.

The difference in values of κ for different priors when α is fixed illustrates the tension between Bayesian and classical analysis. In the next subsection the relation between κ and α is analyzed further. Note again that this relation is independent from data that are actually

observed. The consequence of a sensitivity analysis for the prior that incorporates both κ and α as valuable inputs is a choice of an α-level. This can be used to judge the p-value of the data at hand: $p = P_{y^*|H_0}(\mathrm{MLR}_{y^*}(\widehat{\theta}_{\mathrm{ML}}) > \mathrm{MLR}_{\underline{y}^*}(\widehat{\theta}_{\mathrm{ML}}))$, where $\mathrm{MLR}_{\underline{y}^*}$ is the observed likelihood ratio.

3.4.3 The relation between prior, κ, α and n

The distribution of the marginal likelihood ratio has a limiting distribution under the null hypothesis in terms of $\gamma = n(1-\rho)$, as provided in section 2.3.3. This asymptotic distribution gives a remarkable good approximation in finite samples, even as small as $n = 25$. Consequently, priors formulated in terms of γ imply an almost fixed relation between κ and α values for different values of n. The relation in figure 3.1 for $n = 100$ is almost indistinguishable from that obtained for $n = 1{,}000$, when the priors $\pi_1(\gamma) \sim U(0; 50)$, and $\pi_2(\gamma) \sim 1/14.47 \exp(-\gamma/15)$ are used. For $n = 100$ these priors correspond to the priors (3.30) and (3.31).

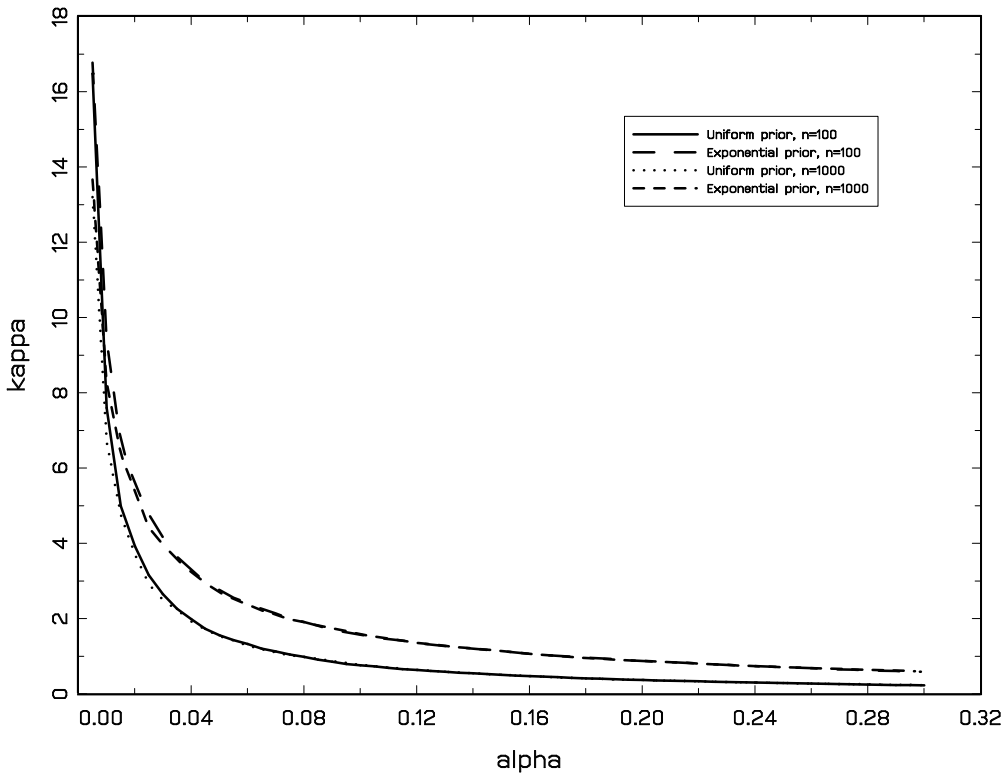

Figure 3.1: The relation between κ and α values.

The relation between κ and α strongly depends on the prior $\pi(\gamma)$. In section 3.3 it was deduced that $\kappa = \int f(\theta, \psi_\alpha)\pi(\theta|H_1)\mathrm{d}\theta$, where ψ_α is the α critical value of S under H_0. In the AR(1) model where $\theta := \gamma$ and $S := \widehat{\gamma}_{\mathrm{ML}}$, the relation holds approximately. The use of this relation is illustrated by figure 3.2, where the α-quantiles are given for the asymptotic

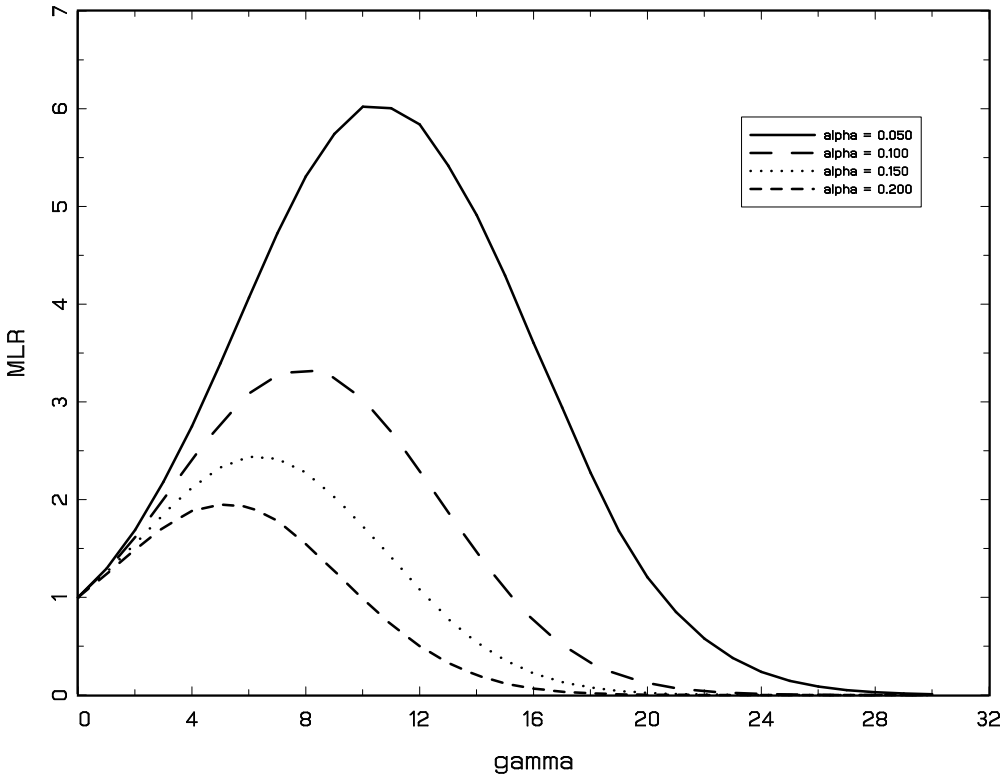

Figure 3.2: Quantiles of the marginal likelihood ratio of the AR(1) model.

distribution of the marginal likelihood ratio, as a function of γ. Let us consider the 5% quantile function to explain the figure. This function has its maximum in $\gamma = 11$, and a maximum marginal likelihood ratio of about 6, corresponding to a marginal loglikelihood of 1.8. These values coincide with the 5% critical values as provided in table 2.2. The α-quantile functions provide all necessary information to compute the value of κ corresponding to α for any $\pi(\gamma)$. An interesting example is provided by uniform priors $\gamma \sim U(0; K)$ with $K \geqslant 30$. For a uniform prior with $K = 30$ it can be calculated that a 5 percent size is obtained for $\kappa = 2.6$. As for $K > 30$ the α-quantile is virtually zero, this size is obtained for

$$\kappa = 2.6 \times \frac{30}{K}. \tag{3.32}$$

Note that for $K = 50$ a 5 percent size is obtained for $\kappa = 1.56$, corresponding to the value in table 3.2. In general for other priors it is impossible to provide an analytical expression for the relation between κ and K. Low values of κ corresponding to $\alpha = 0.05$ are obtained for priors with much probability mass for $\gamma > 25$ and/or near $\gamma = 0$. For this reason noninformative priors which are infinite at $\gamma = 0$ seem to be less appropriate.

Most priors are formulated in terms of ρ instead of γ. This has interesting consequences. The relation between κ and α then depends on n. This explains Andrews' statement that his

Table 3.3: The size for different number of observations.

n	$\pi_1(\rho)$	$\pi_2(\rho)$
25	0.171	0.032
50	0.097	0.060
100	0.050	0.050
250	0.020	0.028
1000	0.004	0.007

asymptotic results do not apply to tests of unit roots. A clear illustration of this dependence on n is given by the uniform prior $\rho \sim U(0.5; 1)$. In terms of $\gamma = n(1-\rho)$ this prior corresponds to $\gamma \sim U(0; 50)$ for $n = 100$, and for $n = 1{,}000$ to $\gamma \sim U(0; 500)$. For $n = 100$ this prior implies that $\kappa = 1.56$ corresponds to $\alpha = 0.05$, and for $n = 1{,}000$, κ follows from (3.32), so $\kappa = 0.156$. Table 3.3 gives the implied size as a function of n for the priors $\pi_1(\rho)$, and $\pi_2(\rho)$ with κ chosen such that for $n = 100$, the size is 5 percent.

Thus, if one specifies uniform priors in terms of ρ, the size of the corresponding test is of order α/n (as table 3.3 shows this is less clear for the exponential prior). Whether this is a desirable thing, is a matter of taste. The intuitive notion that nonstationarity should show in the long run is confirmed if the prior is formulated in terms of ρ, which seems to be the most natural thing to do.

3.5 Conclusion

In this chapter we compared classical and Bayesian inference on a single covariance parameter in the general linear model, where β and σ^2 are regarded as nuisance parameters. The independence Jeffreys' prior implies duality between classical and Bayesian marginal likelihood. In the case that the marginal likelihood contains only one parameter and the ratio is monotone in some statistic, classical marginal likelihood ratio and Bayesian posterior odds tests use the data in the same way.

We applied these results on the regression model with first order autoregressive disturbances, specified in the unobserved component format. Using the independence Jeffreys' prior provides a Bayesian marginal likelihood that is proportional to the classical marginal likelihood. As the marginal likelihood ratio is approximately monotone in $\hat{\rho}$, the classical and Bayesian tests are undistinguishable. The power functions coincide and are very close to the power envelope.

The only relevant discussion between classical and Bayesian statisticians is the proper size of the tests. Whether this is a fruitful discussion remains to be seen. The influence of the prior is very strong, the effect of the sample size is a complicating factor. Reference priors are ambiguous and will also give different results when γ is used instead of ρ. The result (3.21) on "weighted quantiles of the marginal likelihood ratio" provides a direct way to understand the role of the prior.

Chapter 4

Marginal likelihood in state-space models

Abstract

This chapter deals with inference on parameters in the system matrices of state-space models with diffuse initial conditions. The diffuse Kalman filter can produce both the profile and diffuse likelihood, depending on whether the initial condition is treated as a fixed or random variable. In literature on Kalman filtering almost no motivation is provided what likelihood to use.

We argue that the marginal likelihood and consequently – with some care – the diffuse likelihood, is to be preferred for estimation and testing. We provide simple adjustments needed in the diffuse Kalman filter to obtain the marginal likelihood.

The diffuse likelihood depends on the specific state-space representation of the model, leading to problems in some situations. These problems do not occur in the marginal likelihood.

Formally in nonlinear models the diffuse and marginal likelihood may not be used for inference, because these likelihoods are parameter dependent transformations of the data. An alternative estimation method is provided, based on a first order approximation.

When dealing with competing models, the marginal and diffuse likelihood cannot be used for goodness of fit measures, such as the Akaike information and Bayesian information criterion, because for different models these likelihoods are based on different transformations of the data. For nested models an alternative procedure is provided based on the marginal likelihood.

4.1 Introduction

State-space models with unknown initial conditions arise in time series models containing regression parameters, time-varying parameters, and nonstationary components. The diffuse Kalman filter and the exact initial Kalman filter provide efficient ways to cope with the situation of an unknown initial condition. The filters produce a profile or diffuse likelihood, depending on whether the initial condition is treated as a fixed or random variable. The likelihood is used for inference on parameters of the covariance structure, i.e. the parameters in the system

matrices. In literature on Kalman filtering almost no motivation is provided what likelihood to use. There are a few exceptions. Shephard and Harvey (1990) compare diffuse and profile likelihood based inference on the signal-to-noise ratio in the local level model and advise to use the diffuse likelihood. Shephard (1993) and Kuo (1999) advocate to use the marginal likelihood in a regression model with a stochastic trend component and the diffuse filter to compute it. We argue too that the marginal likelihood and consequently – with some care – the diffuse likelihood, is to be preferred for estimation and testing.

In terms of the general linear model, the marginal likelihood is the likelihood of a transformation of the observations y such that the transformed data is independent of the location parameters β and the scale parameter σ, the nuisance parameters. In literature on marginal likelihood a number of examples is provided where inference based on marginal likelihood is superior to that based on profile likelihood, see for example Cooper and Thompson (1977), Tunnicliffe Wilson (1989). In the standard profile likelihood approach the nuisance parameters are being replaced by appropriate maximum likelihood estimates. This approach can result in biased estimates of θ and test procedures with disappointing small sample properties. The use of marginal likelihood helps to reduce bias and to improve power of tests, especially in small samples, see Rahman and King (1997) and chapter 2.

There is a difference between the diffuse and marginal likelihood. Unlike the marginal likelihood, the diffuse likelihood depends on the specific state-space representation of the model, causing problems in some situations. However, it is easy to derive the marginal likelihood from the diffuse likelihood. Actually, the most convenient way to compute the marginal likelihood is to use the diffuse Kalman filter. Only minor modifications of the diffuse Kalman filter recursions are needed in order to get the marginal likelihood.

In section 4.2 the state-space model is defined and different likelihood concepts are discussed. In the subsequent sections the differences are illustrated in examples concerning parameter estimating and testing, and model comparison. Section 4.3 concerns the difference between the marginal and the profile likelihood. As an example the linear model with first order autoregressive disturbances, as given in chapter 2, is discussed. Section 4.4 provides a simple model where different state-space representations, result in a parameter dependent change of the diffuse likelihood, while the marginal likelihood remains the same. Inference in nonlinear models is discussed in section 4.5 . Formally in nonlinear models the diffuse and marginal likelihood may not be used for inference, because these likelihoods are parameter dependent transformations of the data. An alternative iterative estimation method is provided, based on a first order approximation, circumventing the problem of the parameter dependent transformation of the data. Section 4.6 treat the topic of model comparison. When dealing with competing models, the marginal and diffuse likelihood cannot be used for goodness of fit measures, such as the Akaike information and Bayesian information criterion, because for different models these likelihoods are based on different transformations of the data. For nested models an alternative procedure for testing regression parameters is provided. Section 4.7 concludes.

4.2 Different likelihood concepts in the state-space model with diffuse initial condition

The state-space model with unknown initial conditions is provided by:

$$\begin{array}{lcllcll} y_t & = & Z_t\alpha_t + \varepsilon_t, & \varepsilon_t & \sim & N(0,\sigma^2 H_t), & \\ \alpha_{t+1} & = & T_t\alpha_t + R_t\eta_t, & \eta_t & \sim & N(0,\sigma^2 Q_t), & t = 1,\ldots,T, \\ \alpha_1 & = & a_0 + A_0\delta + R_0\eta_0, & \eta_0 & \sim & N(0,\sigma^2 Q_0), & \\ \delta & \sim & N(\delta_0,\sigma^2\Sigma), & & & & \end{array} \tag{4.1}$$

where y_t is an $(n_t \times 1)$ vector of observations and α_t is an unobserved state vector. We assume that ε_t, η_t, and δ are uncorrelated. The $(d \times 1)$ vector δ deals with the initial condition. $\Sigma = 0$ corresponds to an unknown fixed initial condition, and $\Sigma^{-1} = 0$ corresponds to a diffuse initial condition.

Define ψ as the parameters in the system matrices Z_t and T_t, and θ as the parameters in the matrices R_t, H_t, and Q_t concerning the covariance structure. This state-space model can be expressed as a general linear model with some covariance structure,

$$y = X\beta + u,\ u \sim N(0,\sigma^2\Omega),\ \Omega = \Omega(\theta), \tag{4.2}$$

where $y = \left[\ y_1' \cdots y_T'\ \right]'$, and X is an $(n \times d)$ matrix provided by

$$X = \left[\ (Z_1A_0)' \quad (Z_2T_1A_0)' \cdots \left(Z_T\Pi_{t=T-1}^{1}T_tA_0\right)'\ \right]',$$

and $n = \sum_{t=1}^{T} n_t$. We apply the usual notation for the general linear model and use the $(k \times 1)$ vector β instead of the $(d \times 1)$ vector δ. Ω has a complicated,but specified structure as induced by the state-space model. Without loss of generality it is assumed that $a_0 = 0$. Note that the matrix X may depend on ψ.

The likelihood for the state-space model (4.1) can be evaluated by the Kalman filter. De Jong (1991a) and De Jong (1991b) provides an augmented version of the Kalman filter, the diffuse Kalman filter, and Koopman (1997) provides an exact initial Kalman filter to deal with the unknown initial condition.

For $\Sigma = 0$, the (β,σ)-maximized loglikelihood in terms of the general linear model (4.2) is provided by

$$-2\ell_P(\theta) = n\left(1 + \ln 2\pi + \ln y'\Omega^{-1}M_X^{\Omega}y - \ln(n)\right) + \ln|\Omega|, \tag{4.3}$$

where $M_X^{\Omega} = I - X(X'\Omega^{-1}X)^{-1}X'\Omega^{-1}$. We refer to (4.3) as the profile loglikelihood.

For $\Sigma^{-1} = 0$ the loglikelihood $\ln f(y) = \ln f(\beta) + \ln f(y|\beta) - \ln f(\beta|y)$ does not exist. De Jong (1991a) defines for $\Sigma^{-1} \to 0$ the diffuse loglikelihood as $\ln f(y) + \frac{1}{2}\ln|\Sigma|$. In De Jong and Chu-

Chun Lin (1994) the definition is slightly different, $\ln f(y) + \frac{1}{2}\ln|\sigma^2\Sigma|$, resulting in a smaller number of degrees of freedoms. In terms of the general linear model the latter, concentrated with respect to σ^2, is provided by

$$-2\ell_D(\theta, \widehat{\sigma}^2_{\mathrm{ML}}) = m\left(1 + \ln 2\pi + \ln y'\Omega^{-1}M_X^{\Omega}y - \ln(m)\right) + \ln|\Omega| + \ln\left|X'\Omega^{-1}X\right|, \tag{4.4}$$

where $m = (n-k)$. Formally the derivation of the diffuse likelihood is Bayesian as β is treated as a random variable.

An alternative to the diffuse likelihood is the marginal likelihood. The concept of marginal likelihood was introduced in chapter 2. The marginal likelihood with respect to β, denoted by $L_{M_\beta}(\theta, \sigma^2)$ is given in (2.6). For inference on θ the scale parameter σ is still a nuisance parameter. σ can be integrated out of the likelihood, resulting in a concentrated likelihood, indicated by $\widehat{\sigma}^2_{\mathrm{ML}}$. An alternative is the marginal likelihood with respect to β and σ, denoted by $L_{M_{\beta,\sigma}}(\theta)$, as provided in (2.8). In appendix A.4 recursions are provided for calculating the marginal likelihood in state-space-models.

Table 4.1: Differences between different likelihood concepts, apart from constants.

Δ	-2Δ
$\ell_D(\theta, \widehat{\sigma}^2_{\mathrm{ML}}) - \ell_P(\theta)$	$-k\ln y'\Omega^{-1}M_X^{\Omega}y + \ln\lvert X'\Omega^{-1}X\rvert$
$\ell_{M_\beta}(\theta, \widehat{\sigma}^2_{\mathrm{ML}}) - \ell_D(\theta, \widehat{\sigma}^2_{\mathrm{ML}})$	$-\ln\lvert X'X\rvert$
$\ell_{M_{\beta,\sigma}}(\theta) - \ell_{M_\beta}(\theta, \widehat{\sigma}^2_{\mathrm{ML}})$	$-m\ln y'M_X y$
ℓ_P	Profile loglikelihood
ℓ_D	Diffuse loglikelihood
ℓ_{M_β}	Marginal loglikelihood with respect to β
$\ell_{M_{\beta,\sigma}}$	Marginal loglikelihood with respect to β and σ

Table 4.1 provides an overview of the differences between the likelihood concepts. Let us first assume that X does not depend on ψ. In that case the diffuse and (concentrated) marginal likelihood are proportional, so their differences are not relevant for inference on θ. The main difference with the profile likelihood is the term $\widehat{\sigma}^2_{\mathrm{ML}}\left(X'\Omega^{-1}X\right)^{-1}$, the covariance matrix of $\widehat{\beta} = \left(X'\Omega^{-1}X\right)^{-1}X'\Omega^{-1}y$. Estimation based on the marginal (and so the diffuse) likelihood is to be preferred since it adjusts for the evidence on Ω in the part of the data that is a linear function of X, which is pseudo-information. The case with k observations is insightful since Ω drops out of the marginal likelihood. An extreme example of the difference between the profile and marginal likelihood was provided in chapter 2 for the linear model with AR(1) disturbances. Section 4.3 provides more details.

In the more general case that the matrix X depends on ψ, the differences between the diffuse and (concentrated) marginal likelihood are relevant for inference on θ and ψ. Let us first assume that X depends on ψ in a linear way, such that $X = X^*D$, with D an $(k \times k)$ nonsingular matrix that depends on ψ, and X^* is independent of ψ. In this case the concentrated marginal

likelihood ℓ_{M_β} and the marginal likelihood $\ell_{M_{\beta,\sigma}}$ are proportional, because the term $y'M_Xy$ does not depend on ψ. The difference between the marginal and diffuse likelihood is the term $|X'X|$. The marginal likelihood is to be preferred as it is invariant to regular transformations of X. In section 4.4 an example is provided for a simple model where the diffuse likelihood is sensitive to different state-space representations, while the marginal likelihood is not.

When X depends on ψ in a nonlinear way, the diffuse likelihood cannot be used for inference on θ and ψ, as the term $|X'X|$ depends on ψ. In this situation also the difference between $\ell_{M_{\beta,\sigma}}$ and ℓ_{M_β} becomes relevant, because $y'M_Xy$ depends on ψ. Formally the marginal likelihood may not be used for inference on θ and ψ, because for each value of ψ the likelihood of a different ψ-dependent transformation of y is determined. In practice the marginal likelihood seems to be a reasonable choice. In section 4.5 an alternative estimation procedure based on a first order approximation is provided.

4.3 Profile and marginal likelihood

In general inference on θ by marginal likelihood is preferable to profile likelihood, see for example King (1980) and Rahman and King (1997). Unlike the profile likelihood, the marginal (and diffuse) likelihood has zero expectation of the score function as the likelihood is based on the density of a random variable and hence can give unbiased estimates of θ, see Shephard (1993) and Kuo (1999). They want to estimate the signal-to-noise ratio in a stochastic trend components model. The profile likelihood produces many zero estimates when in the data generating process this ratio is positive. The marginal likelihood reduces the problem of zero estimates of signal-to-noise ratios, see also Shephard and Harvey (1990).

An extreme example of the difference between the marginal (and diffuse) versus the profile likelihood was provided by the famous econometric "unit root problem" in chapter 2. The core of this problem is that the profile likelihood degenerates in the unit root. The marginal likelihood for the model specified as (2.1)–(2.3) is well-defined for $-1 < \rho \leqslant 1$, see section 2.2.2. In section 2.3.4 it was shown that marginal likelihood ratio tests on unit root outperform other tests known from literature, especially in small samples.

The computation of the marginal likelihood by the diffuse Kalman filter in the unit root case is not straightforward. Actually, (2.1)–(2.3) consists of two different models, and there is not a state-space representation that is valid for both $\rho = 1$ and $|\rho| < 1$, leading to computational problems when $\rho = 1$. An alternative to circumvent this problem, is to restate (2.1)–(2.3) in first differences. This requires the computation of the covariance matrix of the initial state. For simple models this can be done analytically, see the the ARMA(1,1) example in section 2.4.2. The state-space formulation for this model in first differences is given by (2.25)–(2.27). From (2.27) it follows that the initial condition is well-defined for $\rho = 1$, and the diffuse Kalman filter can be used to compute the marginal likelihood.

For more complex models it might be difficult to derive the unconditional variance of α_t. A simple alternative is to use the basic model (2.1)–(2.3) in levels for $|\rho| < 1$ and to compute the marginal likelihood in $\rho = 1$ by approximation in $\rho = 1 - \varepsilon$, e.g. 0.999. More elegant is to use a separate state-space formulation in $\rho = 1$. The resulting marginal (and likewise diffuse) likelihood can be compared and used for inference as they are based on the same transformation of y.

4.4 Diffuse and marginal likelihood

The difference between the diffuse and marginal likelihood is the determinant of $X'X$. In terms of the state-space model (4.1) $X'X$ can be expressed as

$$X'X = A_0' \sum_{t=1}^{T} \left[\left(\prod_{i=1}^{t-1} T_i' \right) Z_t' Z_t \left(\prod_{i=t-1}^{1} T_i \right) \right] A_0,$$

which may depend on ψ, the unknown parameters in the system matrices T_t and Z_t. If this is the case, inference on ψ based on the diffuse likelihood will differ from that based on the marginal likelihood.

The marginal likelihood is to be preferred as it is invariant to regular transformations of X. In the case $X = X^*D$, with D an $(k \times k)$ nonsingular matrix that depends on ψ and X^* is independent of ψ, the diffuse likelihood cannot be used for inference on θ and ψ. This is illustrated by the following example.

Consider the model

$$y_{jt} = \mu_j + \phi_j \lambda_t,\ \lambda_{t+1} = \lambda_t + \eta_t, \tag{4.5}$$

for $j = 1, 2$, $t = 1, \ldots, T$, and $\phi_1 = 1$. For this simple model two different state-space formulations are provided, resulting in two different diffuse likelihoods, but the same marginal likelihood. The difference in the two state-space formulations concerns the initial condition. As model (4.5) is not identified an extra restriction has to be imposed. In the first state-space formulation this restriction is $\lambda_1 = 0$, and in the second $\mu_2 = 0$, corresponding with $A_0 = \begin{pmatrix} \mathbf{0}_2 & I_2 \end{pmatrix}'$, and $A_0 = \begin{pmatrix} I_2 & \mathbf{0}_2 \end{pmatrix}'$, respectively. For both specifications it holds that $\alpha_t = \begin{pmatrix} \lambda_t & \mu_t' \end{pmatrix}'$, $Z_t = \begin{pmatrix} \phi & I_2 \end{pmatrix}$, $T = I_3$, $H_t = 0$, $R_t = \begin{pmatrix} 1 & \mathbf{0}_2' \end{pmatrix}'$, $R_0 = 0$, and $Q_t = 1$. Let X_i denote the matrix X in specification i, then $|X_1'X_1| = T^2$, $|X_2'X_2| = T^2\phi_2^2$, and $X_2 = X_1 D$, where $D = \begin{pmatrix} 1 & 1 \\ \phi_2 & 0 \end{pmatrix}$.

The difference in the diffuse loglikelihood between the two specifications is $\ln \phi_2$. Only in the first specification inference on ϕ based on the diffuse and marginal likelihood leads to the same results. The marginal likelihood does not depend on the specification.

4.5 Nonlinear state-space models

In this section we consider the case that in the general linear model (4.2) the matrix X depends on ψ in a nonlinear way, for example $X = X^{\psi}$. In this context we define nonlinear as: there is no regular transformation of $X = X(\psi)$, such that $X = X^*D$, where D is an $(k \times k)$ nonsingular matrix that depends on ψ, and X^* is independent of ψ.

Two different approaches are provided for inference on θ and ψ. The first approach is based on maximization of the likelihood function with respect to θ and ψ. The second approach is an iterative procedure based on a linear approximation of the model.

In the first approach the differences between the likelihoods, as provided in table 4.1, matter because they all depend on θ and ψ. From the previous section it is clear that the diffuse likelihood cannot be used for inference on θ and ψ, because the diffuse likelihood is even sensitive to regular transformations of X.

One has to be cautious to use the marginal likelihood, because it is a likelihood of a ψ dependent transformation of y, resulting in a Jacobian $\widehat{\sigma}_{\mathrm{ML}}^{k} \left(|X'X| \,/\, |X'\Omega^{-1}X| \right)^{1/2}$, see section 2.2.1.

For the special case that $\Omega = I$ the Jacobian is $\widehat{\sigma}_{\mathrm{ML}}^{k}$ and the marginal loglikelihood $\ell_{M_\beta}(\theta, \widehat{\sigma}^2_{\mathrm{ML}})$ is the profile loglikelihood, corrected for degrees of freedom. Note that in this case the marginal loglikelihood $\ell_{M_{\beta,\sigma}}(\theta)$ is zero, thus cannot be used for inference on ψ. In case of a covariance structure the ratio of these determinants, conditional on θ is in practice not very sensitive to ψ. This is illustrated by the next example.

Consider the state-space model

$$y_t = \alpha_t + \mu + t^{\psi}\beta, \;\; \alpha_{t+1} = \rho\alpha_t + \eta_t, \tag{4.6}$$

where $\alpha_1 \sim N(0, \sigma^2/(1-\rho^2))$, $-1 < \rho < 1$, and $\psi > 0$. In figure 4.1 a typical example of the different likelihood functions are provided. Data is generated from (4.6), with $\rho = 0.5$ and $\psi = 0.5$ ($\mu = 10$, $\beta = 1$, $\sigma^2_\eta = 1$). Unlike the marginal and profile likelihood, the diffuse likelihood function has no optimum: for $\psi \to 0$ the diffuse likelihood goes to infinity.

If in (4.6) t^{ψ} is replaced by $\left(t^{\psi} - 1\right)/\psi$, the marginal likelihood remains the same, because it is a regular transformation of the original model. However, the diffuse likelihood function changes and almost coincides with the marginal likelihood. As expected, the profile loglikelihood goes to minus infinity as $\rho \uparrow 1$. This corresponds to the fact that in a simulation study underestimation of ρ is the main problem of the profile likelihood.

The second approach for inference on θ and ψ is to linearize the model $y = X_1\beta_1 + X_2(\psi)\beta_2 + \varepsilon$. Let ψ_0 be a trial value of ψ. Expanding about ψ_0 gives approximately

$$y = X_1\beta_1 + X_2(\psi_0)\beta_2 + x_2(\psi_0)\gamma + \varepsilon, \tag{4.7}$$

where $\gamma = (\psi - \psi_0)\beta_2$, and $x_2(\psi_0) = \partial X_2(\psi)/\partial\psi$ evaluated in ψ_0.

The linearized model can be used for likelihood inference on θ. Conditional on ψ_0 the diffuse likelihood $\ell_D(\theta, \widehat{\sigma}^2_{\mathrm{ML}})$ and the marginal likelihoods $\ell_{M_\beta}(\theta, \widehat{\sigma}^2_{\mathrm{ML}})$ and $\ell_{M_{\beta,\sigma}}(\theta)$ are proportional, so it makes no difference which likelihood is used.

A new value for ψ can be obtained from previous estimates of β_2 and γ:

$$\psi^{(i+1)} = \psi^{(i)} + \widehat{\gamma}^{(i)}/\widehat{\beta}_2^{(i)}, \tag{4.8}$$

where (i) denotes iteration step i. Conditional on ψ_2 a new estimate of θ and σ^2 is obtained by maximizing the likelihood function. When the iteration converges we have $\psi^{(i+1)} = \psi_i^{(i)}$ or $\widehat{\gamma}^{(i)} = 0$. For more details on Gauss Newton Regression, see for example Davidson and MacKinnon (1993).

For the profile likelihood both approaches provide the same results. This is not the case for the marginal likelihood, because 1) the determinant terms in the nonlinear model and the linearized model are different and 2) in the first approach it is assumed that $\partial \ell_{M_\beta}(\theta, \sigma^2, \psi)/\partial\psi = 0$, while in the second approach this restriction is replaced by $\widehat{\gamma} = 0$. In practice the differences are expected to be small.

In state-space models where in the observation equation y_t depends in a nonlinear way on the state vector α_t, and in the transition equation α_{t+1} depends on α_t in a nonlinear way, only the second approach is applicable, see Durbin and Koopman (2001, ch. 11).

4.6 Testing in nested models

In this section we deal with competing models. Well known and often used likelihood based tests are the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). In these measures the loglikelihood, evaluated in the estimated parameters, is corrected for the number of parameters in the model in order to have a fair comparison. It makes no sense to use the diffuse likelihood in AIC and BIC measures, as is proposed by Durbin and Koopman (2001), because the diffuse likelihood depends on the specific state-space formulation (see section 4.2), and so the same holds for AIC and BIC. Marginal likelihood is based on a transformation of the data. Different models imply different transformations, so marginal likelihood cannot be used for AIC and BIC measures either.

When dealing with nested models alternative tests are available. Consider the case that we want to test the null hypothesis H_0 that $\beta_2 = 0$, against the alternative hypothesis H_1 that $\beta_2 \neq 0$ in $y = X\beta + \varepsilon$, where $X = \begin{bmatrix} X_1 & X_2 \end{bmatrix}$ is an $(n \times k)$ matrix and $\beta = \begin{pmatrix} \beta_1' & \beta_2 \end{pmatrix}'$. We only treat the case that X_2 is a single variable, but this may easily be extended to the case that X_2 contains more than one variable. Using the modified diffuse Kalman filter, one obtains not only the marginal likelihood $\ell_{M_{\beta,\sigma}}(\theta)$, but also $\widehat{\beta} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y$, conditional on θ. Running the two filters including and excluding X_2 would provide two marginal likelihoods of different dimensions. The solution is to run the modified diffuse Kalman filter including X_2

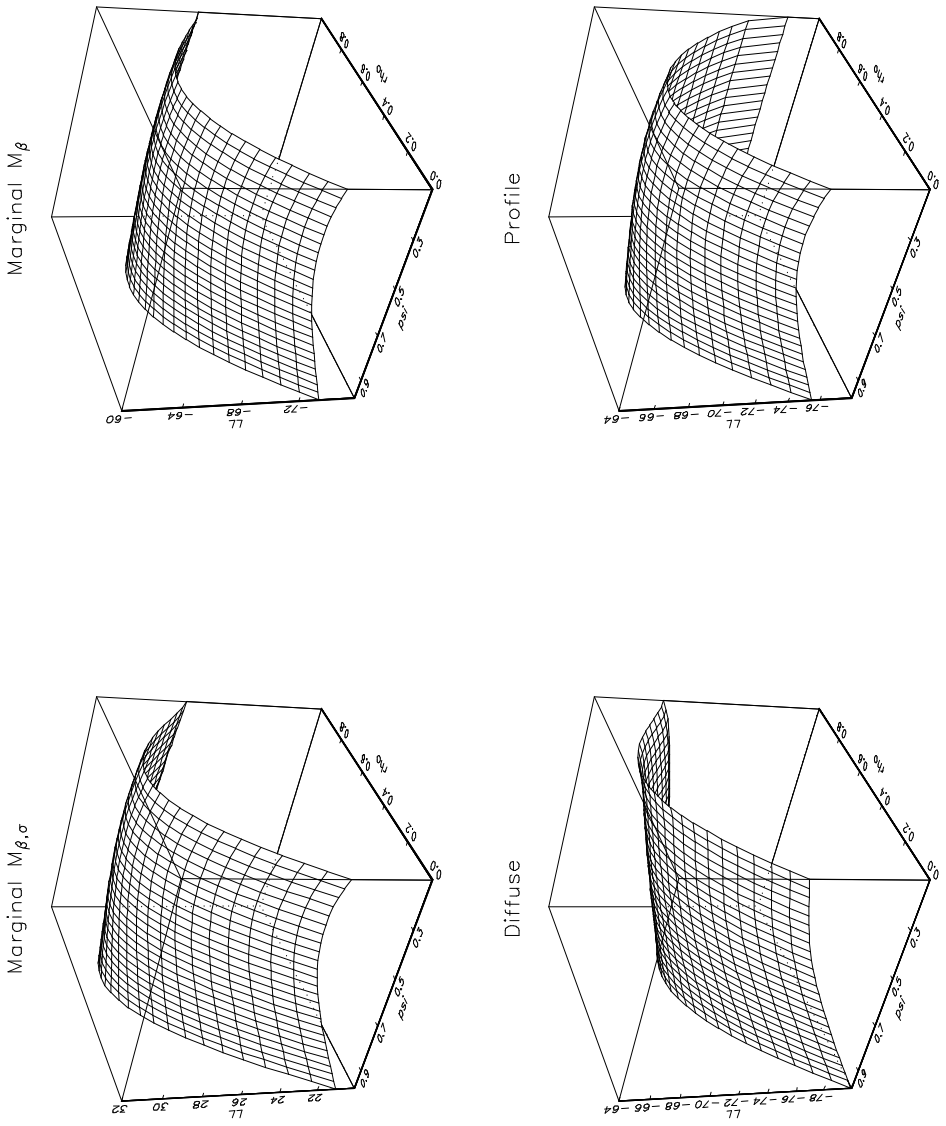

Figure 4.1: Different likelihoods as a function of ρ and ψ in the nonlinear model.

and to consider $\ell_{M_{\beta,\sigma}}(\theta)$ and $\widehat{\beta}$ conditional on θ simultaneously.

A classical approach would be to compute $\widehat{\theta}_{\mathrm{ML}}$ and to use the t-value of $\widehat{\beta}_2 \mid \widehat{\theta}_{\mathrm{ML}}$ for testing. The test T_3 is based on the marginal likelihood, so $\widehat{\theta}_{\mathrm{ML}} = \arg\max \ell_{M_{\beta,\sigma}}(\theta)$, and test T_4 is based on the profile likelihood, so $\widehat{\theta}_{\mathrm{ML}} = \arg\max \ell_P(\theta, \widehat{\sigma}^2_{\mathrm{ML}})$, where $\widehat{\sigma}^2_{\mathrm{ML}} = y'\Omega^{-1}M^{\Omega}_X y/m$. Conditional on the maximum likelihood estimate the t-value has a t-distribution.

A Bayesian criticism on this approach is that the uncertainty in $\widehat{\beta}_2$ is underestimated. Bayesians would prefer to use a prior $\pi(\theta)$ and to compute

$$f(\beta_2|y) = \int f(\theta, \beta_2|y)d\theta = \int f(\beta_2|\theta, y)f(\theta|y)d\theta, \tag{4.9}$$

where $f(\theta|y) \propto f(y|\theta)\pi(\theta)$. It is shown in chapter 3 that the marginal likelihood $\ell_{M_{\beta,\sigma}}(\theta)$ is proportional to $f(y|\theta)$, where

$$f(y|\theta) = f(y|\beta, \sigma^2, \theta)\pi(\beta, \sigma^2|\theta)/f(\beta, \sigma^2|y, \theta), \tag{4.10}$$

when the independence Jeffreys' prior $\pi(\beta, \sigma^2|\theta) \propto \sigma^{-2}\left|X'X\right|^{1/2}$ is used. When assuming a flat prior for θ, the posterior of θ is proportional to the marginal likelihood,

$$f(\theta|y) \propto f(y|\theta) \propto \ell_{M_{\beta,\sigma}}(\theta). \tag{4.11}$$

It can be deduced that $f(\beta|\theta, y)$ has a multivariate t-distribution f_m, see appendix A.5 for a definition. The marginal posterior $f(\beta_2|\theta, y)$ is given by

$$f(\beta_2|\theta, y) \sim f_m(\beta_2|\widehat{\beta}_2, \widehat{\sigma}^2_{\mathrm{ML}}\,(X_2'\Omega^{-1}M^{\Omega}_{X_1}X_2)^{-1}, 1), \tag{4.12}$$

where $\widehat{\beta}_2 = (X_2'\Omega^{-1}M^{\Omega}_{X_1}X_2)^{-1}X_2'\Omega^{-1}M^{\Omega}_{X_1}y$, see for example Poirier (1995, p. 126, 127). The posterior $f(\beta_2|y)$ can be calculated from (4.12) and (4.11) by numerical integration.

We compare the classical tests with "Bayesian tests" based on confidence intervals of the posterior of β_2. Test T_1 use a symmetric interval, with 2,5% at each side, and test T_2 use the 95% Highest Posterior Density interval, following Box and Taio (1973).

We compare the different tests in a model that may be expected to be problematic in terms of distinguishing between deterministic terms (sinuses) and stochastic components (AR(1)), for a small sample size ($T = 50$). The data generating process is

$$\begin{aligned} y_t &= \alpha_t + \mu + \beta_1 \sin(t/2) + \beta_2 \sin(t/3), \\ \alpha_{t+1} &= \rho\alpha_t + \eta_t, \end{aligned}$$

for $t = 1, \ldots, 50$, $-1 < \rho < 1$, where $\eta_t \sim N(0, \sigma^2)$, $-1 < \rho < 1$, $\alpha_1 \sim N(0, \sigma^2/(1-\rho^2))$, $(\mu, \beta_1)' = (10, 1)$, and $\sigma^2 = 1$. Power functions are provided in table 4.2 as a function of β_2 for different values of ρ.

Table 4.2: Power functions for testing in nested models.

	$\rho = .1$				$\rho = .5$				$\rho = .9$			
β_2	T_1	T_2	T_3	T_4	T_1	T_2	T_3	T_4	T_1	T_2	T_3	T_4
0	.0494	.0494	.0618	.0847	.0535	.0534	.0730	.1113	.0708	.0712	.0732	.0938
.05	.0541	.0543	.0666	.0924	.0575	.0580	.0761	.1125	.0727	.0724	.0743	.0951
.10	.0668	.0681	.0841	.1138	.0639	.0654	.0837	.1209	.0741	.0745	.0750	.0974
.15	.0895	.0909	.1121	.1478	.0738	.0746	.0944	.1354	.0762	.0771	.0786	.1032
.20	.1218	.1242	.1525	.1934	.0861	.0879	.1100	.1554	.0816	.0830	.0842	.1098
.30	.2239	.2271	.2618	.3233	.1248	.1272	.1534	.2107	.0983	.1004	.1008	.1317
.40	.3586	.3627	.4128	.4816	.1754	.1784	.2180	.2870	.1242	.1277	.1277	.1590
.50	.5167	.5205	.5729	.648	.2514	.2545	.2963	.3758	.1564	.1595	.1591	.1951
.60	.6734	.6770	.7247	.7845	.3333	.3377	.3877	.4740	.1973	.2002	.1995	.2388
.70	.7993	.8035	.8411	.8861	.4283	.4328	.4844	.5746	.2454	.2494	.2472	.2903

The main lessons are to be learnt from the size distortions. The test based on the profile likelihood, T_4, performs worst. The marginal likelihood based test T_3 performs better, as might be expected. The pseudo-Bayesian tests perform even better. Size distortion of T_3 is due to underestimation of the uncertainty. For T_1 and T_2 and moderate values of ρ, there is hardly size distortion left. Noteworthy is that the HPD test T_1 is slightly better than the symmetric test T_2. For high values of ρ the power drops for all tests: distinction between deterministic and stochastic components becomes hard. The example is constructed to be tough, for more observations and other specifications the differences between the methods will generally be smaller.

4.7 Conclusion

The main conclusion is that the marginal likelihood, efficiently computed with the diffuse Kalman filter or exact initial Kalman filter, is suited for inference in complex state-space models and is superior to the standard profile likelihood methods. We have shown how to adjust the diffuse likelihood in order to obtain the marginal likelihood. We have dealt with some potential problems in using the diffuse and/or marginal likelihood, specifically non-uniqueness of the diffuse likelihood, unit roots, nonlinear models and testing of nested models, and found satisfactorily solutions.

Chapter 5

Efficient computation of hierarchical trends

Abstract[1]

To model a large database containing selling prices for houses, in which local trends, general trends, and specific characteristics play a role, we derived a new procedure to implement a state-space model for repeated measurements. The original model is decomposed into two parts, which are treated differently. The first part is ordinary least squares on data in deviation from means. This step provides a prior for coefficients to be used in the second step, which is a Kalman filter, providing estimates of the trends and the parameters. The procedure exploits and illustrates the Bayesian interpretation of a Kalman filter.

5.1 Introduction

This chapter presents a model for stochastic hierarchical trends. The trends are analyzed on two levels, a combination of trends on a general and on a cluster level. The cluster-level trends are specified as being independent of the general trend. The general trend is specified as a random walk with drift, and the cluster-level trends are specified as random walks. Explanatory variables are specified for individual observations. The model is developed for a large database containing selling prices of houses in different neighborhoods and is operational in Amsterdam for the determination of real-estate taxes. Other applications, for instance, might concern the development of stock-market prices, regional developments of economic issues, and so forth. The proposed model is

$$y_{ijt} = \mu_t + \vartheta_{jt} + x'_{ijt}\beta + \varepsilon_{ijt}, \tag{5.1}$$

$$\mu_{t+1} = \mu_t + \kappa + \eta_t, \tag{5.2}$$

[1]This chapter is based on Francke and de Vos (2000).

and

$$\vartheta_{jt+1} = \vartheta_{jt} + \omega_{jt}, \tag{5.3}$$

for $i = 1, \ldots, n_{jt}, j = 1, \ldots, B$, and $t = 1, \ldots, T$, where y_{ijt} denotes the ith cluster at time t. B is the number of clusters and n_{jt} is the number of observations in cluster j at time t, $n_{jt} > 0$ for at least one t. The general trend and the trend for cluster j are indicated by μ_t and ϑ_{jt}, respectively, and x_{ijt} are explanatory variables with coefficients β. The vector β is fixed. $\varepsilon_{ijt} \sim N(0, \sigma_\varepsilon^2)$, $\eta_t \sim N(0, \sigma_\varepsilon^2 q_1)$, and $\omega_{jt} \sim N(0, \sigma_\varepsilon^2 q_2)$ with q_1 and $q_2 > 0$ and $\mathrm{Cov}(\omega_{jt}, \omega_{kt}) = 0$ except for $j = k$.

The general trend μ_t follows a random walk with drift. A more flexible model for the general trend is

$$\mu_{t+1} = \mu_t + \kappa_t + \eta_t \tag{5.4}$$

and

$$\kappa_{t+1} = \kappa_t + \zeta_t. \tag{5.5}$$

The random walk with drift is chosen because in the example the time period is relatively short, so it is not to be expected that κ_t changes in time. In (5.2), κ can be replaced by $z_t\delta$. In that case, the trend is a combination of a deterministic and a stochastic part. The deterministic part could, for example, represent the influence of interest. A straightforward Kalman filter with a diffuse prior could be used to estimate models of this kind, as was shown by De Jong (1991a), De Jong and Chu-Chun Lin (1994), and Harvey (1989). Due to the structure of repeated measurements, in this case a more sophisticated approach is possible. In this chapter, an alternative procedure is shown based on a Bayesian derivation of the likelihood. The original model is decomposed into a model containing means per time per cluster and a model containing deviations from these means. The model containing deviations from the means does not contain trends. In a first round, a posterior of β, given this information, is obtained from this model by ordinary least squares (OLS) estimation (the priors for β are taken noninformative). The second round is a straightforward Kalman filter starting with this posterior for β as a prior, providing trends and a further update of β. In Section 5.2, this procedure and its derivation are shown. An advantage of this approach is that the number of observations in the Kalman filter is reduced considerably because, at any point in time, for each cluster one sufficient statistic, the mean, replaces n_{jt} observations. Another advantage is that an initialization problem, due to the presence of the explanatory variables x_{ijt} in (5.1), is avoided. This is also shown in Section 5.2.

In the proposed model, it is assumed that the covariances between innovations of the trends are 0. This assumption is not necessary for the method just described. The covariance matrix may have any structure.

For the application presented in this chapter, initialization is complicated by the possibility that in a certain cluster the first observation is not necessarily the first point of the time series.

The choice to start with a large variance matrix fails because of numerical problems. This problem is handled in Section 5.3.

In Section 5.4, an alternative to the classical smoothing algorithm is applied, the forward-backward algorithm, see Merkus, Pollock, and de Vos (1993). This smoothing algorithm can easily be used in cases in which parts of the state vector are not defined as a result of the fact that no observations are available up to time t.

In Section 5.5, results for our example of hierarchical trends, selling prices of apartments in different neighborhoods, are discussed briefly.

5.2 Hierarchical trends

5.2.1 The model

In the model presented in the introduction, the variables μ_t and ϑ_{jt} are not identified. This identification problem can be solved by defining $\delta_{jt} = \mu_t + \vartheta_{jt}$ and $\varphi_{jt} = \eta_t + \omega_{jt}$. An alternative solution is to impose a restriction, such as $\mu_1 = 0$ with probability 1. Statistically both procedures are equivalent, but the latter defines a general trend that cannot be easily estimated differently. With this restriction and κ substituted from the transition equation to the measurement equation, (5.1) to (5.3) are rewritten in state-space format as

$$y_t = \begin{pmatrix} 1_{n_t} & D_t \end{pmatrix} \begin{pmatrix} \mu_t \\ \vartheta_t \end{pmatrix} + 1_{n_t} t\kappa + X_t\beta + \varepsilon_t, \tag{5.6}$$

$$\begin{pmatrix} \mu_{t+1} \\ \vartheta_{t+1} \end{pmatrix} = \begin{pmatrix} \mu_t \\ \vartheta_t \end{pmatrix} + \begin{pmatrix} \eta_t \\ \omega_t \end{pmatrix}, \tag{5.7}$$

and

$$\mu_1 = 0 \text{ with probability 1}, \tag{5.8}$$

with $y_t = (y_{1t}', \ldots, y_{Bt}')'$, $y_{jt} = (y_{1jt}, \ldots, y_{n_{jt}jt})'$, $\varepsilon_t = (\varepsilon_{1t}', \ldots, \varepsilon_{Bt}')'$, $\varepsilon_{jt} = (\varepsilon_{1jt}, \ldots, \varepsilon_{n_{jt}jt})'$, $\vartheta_t = (\vartheta_{1t}, \ldots, \vartheta_{Bt})'$, and $\omega_t = (\omega_{1t}, \ldots, \omega_{Bt})'$. $\varepsilon_t \sim N(0, \sigma_\varepsilon^2 I_{n_t})$ and $\omega_t \sim N(0, \sigma_\varepsilon^2 q_2 I_B)$, with $n_t = \sum_{j=1}^{B} n_{jt}$ the number of observations at time t.
D_t is an $n_t \times B$ dummy matrix,

$$D_t = \begin{pmatrix} 1_{n_{1t}} & & 0 \\ & \ddots & \\ 0 & & 1_{n_{Bt}} \end{pmatrix},$$

and the $n_t \times k$ matrix X_t contains the explanatory variables at time t.

5.2.2 Standard inference

In this model μ, ϑ, κ, β, and $\Psi = \{\sigma_\varepsilon^2, q_1, q_2\}$ need to be estimated. In the case that κ and β are excluded from (5.6), the Kalman filter provides a way to evaluate the likelihood and estimates of $\mu_t|\Psi, Y_t$ and $\vartheta_t|\Psi, Y_t$ with $Y_t = (y_1, \ldots, y_t)$. Usually Ψ is estimated by maximum likelihood (ML). In the case in which the fixed variables κ and β are included in (5.6), there are two options for estimating κ and β: the time-invariant fixed variables κ and β can be treated as fixed or as random (Harvey (1989)).

If κ and β are treated as fixed, the Kalman filter provides estimates of $\mu_t|\Psi, \kappa, \beta, Y_t$ and $\vartheta_t|\Psi, \kappa, \beta, Y_t$. Usually κ, β, and Ψ are estimated by ML. This approach was proposed by Harvey (1989). The ML estimate of κ and β are obtained by using the same Kalman filter for y_t and each column of x_t. After regressing the residuals from the Kalman filter, the ML estimate of κ and β are obtained. Estimates of μ and ϑ are not obtained.

If κ and β become a part of the state vector, κ and β are treated as random (see also Harvey (1989)). A straightforward Kalman filter with a diffuse prior produces a likelihood and gives estimates of $\mu_t|\Psi, Y_t$, $\vartheta_t|\Psi, Y_t$, $\kappa_t|\Psi, Y_t$, and $\beta_t|\Psi, Y_t$.

An alternative to obtain estimates of μ, ϑ, κ, and β is presented by the diffuse Kalman filter of De Jong (1991a) and De Jong and Chu-Chun Lin (1994). In this filter, κ and β are not inserted in the state vector and κ and β can be treated both as random and fixed.

The alternative algorithm that we present is much faster if, as in our case, many repeated measurements occur. Moreover, the decomposition into an OLS step - also suited for model exploration - and a step estimating the state vector provides additional understanding.

5.2.3 An example

To understand the basic ideas, consider for a moment the model $y = X\beta + \varepsilon$ with the $k \times 1$ vector β fixed, $\varepsilon \sim N(0, \sigma^2 I)$, and divide y into two independent parts y_1 and y_2 with n_1 and n_2 elements, respectively, and $n = n_1 + n_2$; then $f(y|\beta, \sigma^2) = f(y_1|\beta, \sigma^2) f(y_2|\beta, \sigma^2)$.

In a Bayesian setup, inference on β can be done in two parts, using $f(\beta|y_1, y_2, \sigma^2) \propto f(\beta|y_1, \sigma^2) f(\beta|y_2, \sigma^2)$. With a noninformative prior for the fixed β, the first part gives $f(\beta|y_1, \sigma^2) \propto f(y_1|\beta, \sigma^2)$. The computation of this can also be done by OLS, so $\beta|y_1, \sigma^2 \sim N((X_1'X_1)^{-1}X_1'y_1, \sigma^2(X_1'X_1)^{-1})$.

The second part, $f(y_2|\beta, \sigma^2)$, can be estimated by recursively updating every datapoint by Bayes formulas,

$$f(\beta|y_{2,1}, \ldots, y_{2,t+1}, \sigma^2) \propto f(\beta|y_{2,1}, \ldots, y_{2,t}, \sigma^2) f(y_{2,t+1}|\beta, \sigma^2).$$

This is equivalent to running the Kalman filter for the following state-space model with initial

state the expectation and variance of $\beta_0 = \beta|y_1, \sigma^2$:

$$y_2 = X_{2,t}\beta_t + \varepsilon_{2,t} \tag{5.9}$$

and

$$\beta_{t+1} = \beta_t + \xi_t, \tag{5.10}$$

for $t = 1, \ldots, n_2$, with $\sigma_\xi^2 = 0$ implying that β is fixed (in classical terms).

Obviously, in this model the recursive update in the second part is not very efficient: The final answer, $f(\beta|y_1, y_2, \sigma^2)$, can be obtained directly by OLS. The conclusion is that, as far as a model contains a subset of information with fixed parameters, inference conditional on this subset may best be done by OLS. Thus, in the model with $\sigma_\xi^2 > 0$ in (5.10) so that there is a changing β_t in the second part, an efficient Kalman filter would use OLS for the first period.

Note that, in smoothing, the subset estimated by OLS is irrelevant. If in (5.10) $\sigma_\xi^2 > 0$, the Kalman smoother obtains recursively $f(\beta_t|y_1, y_2)$ from the last observation of y_2 down to the first observation of y_2. Once one has obtained $f(\beta_0|y_{1,}y_2)$, further smoothing is irrelevant, so also, in smoothing, the first part needs no evaluation by a filter.

For inference on σ^2, the Kalman filter evaluates the "likelihood." The standard way in this model is the "big κ" method: Starting with a prior κI_k with $\kappa \to \infty$, the filter is applied and the prediction errors are incorporated from observation $k+1$ (supposing that $X_{1,k}$, the matrix depending on the first k observations, has full column rank). This comes down to a numerical approximation of the Bayesian predictive likelihood

$$\begin{aligned} &f(y_{1,k+1}, \ldots, y_{1,n_1}, y_2|y_{1,1}, \ldots, y_{1,k}, \sigma^2) \\ &= \left(\left|X_{1,k}'X_{1,k}\right| / |X'X|\right)^{1/2} (2\pi\sigma^2)^{-(n-k)/2} \exp(-\frac{y'M_X y}{2\sigma^2}), \end{aligned} \tag{5.11}$$

with $M_X = I - X(X'X)^{-1}X'$. (An equivalent alternative, by the way, is to do OLS on the first k observations and to use the resulting $f(\beta|y_{1,1}, \ldots, y_{1,k}, \sigma^2)$ as initial state.)

Equation (5.11) is proportional to the concentrated likelihood that may directly be obtained by OLS; the only relevant difference is the power $n-k$ of σ in (5.11), but this is a very plausible modification of the OLS result $|X'X|^{-1/2} (2\pi\sigma^2)^{-n/2} \exp(-y'M_X y/(2\sigma^2))$.

As far as inference on σ^2 is concerned (see De Vos (1998) for the implausible role of the $|X'X|^{-1/2} \left|X_{1,k}'X_{1,k}\right|^{1/2}$in comparing different models), the combination of OLS on the first part and a Kalman filter on the second part is easy. The second part has no starting value problem; $f(y_2|y_1, \sigma^2)$ is well defined. So the total likelihood is

$$f(y_{1,k+1}, \ldots, y_{1,n_1}, y_2|y_{1,1}, \ldots, y_{1,k}, \sigma^2) = f(y_{1,k+1}, \ldots, y_{1,n_1}|y_{1,1}, \ldots, y_{1,k}, \sigma^2) f(y_2|y_1, \sigma^2)$$

and ML estimates or posteriors of σ^2 may directly be obtained. (In the model of this example with "fixed" β, the ML estimate is residual sum of squares divided by $(n-k)$.)

5.2.4 The algorithm for hierarchical trends

The same reasoning is used for the more complex model (5.6) and (5.7). First these equations are written in a Bayesian notation. The transition equation defines the prior $\pi(\mu, \vartheta, \kappa, \beta|\Psi)$. In (5.6) and (5.7) μ, ϑ, κ, and β are assumed to be independent, so

$$\pi(\mu, \vartheta, \kappa, \beta|\Psi) = \pi(\mu|\Psi)\pi(\vartheta|\Psi)\pi(\kappa|\Psi)\pi(\beta|\Psi),$$

where $\pi(\mu|\Psi) = \pi(\mu_1|\Psi)\cdots\pi(\mu_T|\mu_{T-1}, \Psi)$ and $\pi(\vartheta|\Psi) = \pi(\vartheta_1|\Psi)\cdots\pi(\vartheta_T|\vartheta_{T-1}, \Psi)$. $\pi(\mu|\Psi)$, $\pi(\vartheta|\Psi)$, $\pi(\kappa|\Psi)$, and $\pi(\beta|\Psi)$ are in this case a representation of (5.6) and (5.7). The measurement equation provides the likelihood $f(y|\mu, \vartheta, \kappa, \beta, \Psi)$, where $y = \begin{pmatrix} y_1' & \cdots & y_T' \end{pmatrix}'$. The posterior is $f(\mu, \vartheta, \kappa, \beta|y, \Psi)$,

$$f(\mu, \vartheta, \kappa, \beta|y, \Psi) \propto \pi(\mu, \vartheta, \kappa, \beta|\Psi) f(y|\mu, \vartheta, \kappa, \beta, \Psi). \tag{5.12}$$

The problem is that $\pi(\vartheta_1|\Psi)$ and $\pi(\beta|\Psi)$ are not well defined, $\pi(\vartheta_1|\Psi) \sim N(0, C_1)$, and $\pi(\beta|\Psi) \sim N(0, C_2)$ with $C_1^{-1}, C_2^{-1} \to 0$. This causes an initialization problem in the Kalman filter. The initialization of ϑ is standard and will be explained in Section 5.3. For κ, a proper prior is assumed. The initialization problem of β becomes critical if, for a long time L, rank(X_1',X_2',..., $X_L')' < k$, with k the number of explanatory variables. A solution for this problem is provided by the general algorithm for a diffuse prior, devised by De Jong (1988).

The initialization problem for β can be circumvented by taking into consideration that a decomposition may provide a proper prior for β, the prior conditional on the information contained in the deviations from the means. For this reason the likelihood will be rewritten in terms of $\overline{y}_{.t}$ and $\widetilde{y}_{.t}$, with $\overline{y}_{.jt} = \sum_{i=1}^{n_{jt}} y_{ijt}/n_{jt}$, $\widetilde{y}_{ijt} = y_{ijt} - \overline{y}_{.jt}$ and the same for x_t. It can easily be demonstrated that $\sum_{i=1}^{n_{jt}}(y_{ijt} - \mu_t - \vartheta_{jt} - t\kappa - x_{ijt}\beta)^2 = \sum_{i=1}^{n_{jt}}(\widetilde{y}_{ijt} - \widetilde{x}_{ijt}\beta)^2 + n_{jt}(\overline{y}_{.jt} - \mu_t - \vartheta_{jt} - t\kappa - \overline{x}_{.jt}\beta)^2$. So the likelihood can be rewritten in terms of means and deviation from the means,

$$f(y|\mu, \vartheta, \kappa, \beta, \Psi) = f(\overline{y}|\mu, \vartheta, \kappa, \beta, \Psi) f(\widetilde{y}|\beta, \Psi), \tag{5.13}$$

with

$$\begin{aligned}
f(y|\mu, \vartheta, \kappa, \beta, \Psi) &= (2\pi\sigma_\varepsilon^2)^{-n/2} \exp(-\frac{1}{2\sigma_\varepsilon^2} e'e), \\
f(\overline{y}|\mu, \vartheta, \kappa, \beta, \Psi) &= (2\pi\sigma_\varepsilon^2)^{-d/2} |H|^{-1/2} \exp(-\frac{1}{2\sigma_\varepsilon^2} \overline{e}' H^{-1} \overline{e}), \\
e &= y - \begin{pmatrix} 1_n & D \end{pmatrix} \begin{pmatrix} \mu' & \vartheta' \end{pmatrix}' - \mathbf{t}\kappa - X\beta, \\
\overline{e} &= \overline{y} - \begin{pmatrix} 1 & \overline{D} \end{pmatrix} \begin{pmatrix} \mu' & \vartheta' \end{pmatrix}' - \overline{\mathbf{t}}\kappa - \overline{X}\beta,
\end{aligned}$$

$X = \begin{pmatrix} X_1' & \cdots & X_T' \end{pmatrix}'$, $\mathbf{t} = \begin{pmatrix} 1 \times \mathbf{1}_{n_1}', & \cdots & n \times \mathbf{1}_{n_T}' \end{pmatrix}'$, $n = \sum_{t=1}^{T} n_t$, $d = \sum_{t=1}^{T} \sum_{j=1}^{B} 1_{\{n_{jt}>0\}}$,

and $H = \text{diag}\{n_{11}^{-1}, \ldots, n_{Bn}^{-1}\}$; $f(\overline{y}|\mu, \vartheta, \kappa, \beta, \Psi)$ is the density of the cluster means. The matrices $\overline{D}$, $\overline{X}$ and $\overline{\mathbf{t}}$ simply contain cluster means. From (5.13) it follows that $f(\widetilde{y}|\beta, \Psi) = f(y|\mu, \vartheta, \kappa, \beta, \Psi)/f(\overline{y}|\mu, \vartheta, \kappa, \beta, \Psi)$,

so

$$f(\widetilde{y}|\beta, \Psi) = (2\pi\sigma_\varepsilon^2)^{-(n-d)/2} |H|^{1/2} \exp(-\frac{1}{2\sigma_\varepsilon^2}(\widetilde{y} - \widetilde{X}\beta)'(\widetilde{y} - \widetilde{X}\beta)).$$

$f(\widetilde{y}|\beta, \Psi)$ does not depend on $\{q_1, q_2\}$, so $f(\widetilde{y}|\beta, \Psi) = f(\widetilde{y}|\beta, \sigma_\varepsilon^2)$. The decomposition of the likelihood (5.13) is used to rewrite (5.12). The posterior is now given by

$$\begin{aligned} f(\mu, \vartheta, \kappa, \beta|y, \Psi) &\propto f(\overline{y}|\mu, \vartheta, \kappa, \beta, \Psi)\pi(\mu|\Psi)\pi(\vartheta|\Psi)\pi(\kappa|\Psi)\pi(\beta|\Psi)f(\widetilde{y}|\beta, \Psi) \\ &\propto f(\overline{y}|\mu, \vartheta, \kappa, \beta, \Psi)\pi(\mu|\Psi)\pi(\vartheta|\Psi)\pi(\kappa|\Psi)f(\beta|\widetilde{y}, \sigma_\varepsilon^2). \end{aligned} \tag{5.14}$$

For a noninformative prior for β, the posterior is the dual solution to simple regression, so

$$f(\beta|\widetilde{y}, \sigma_\varepsilon^2) \sim N((\widetilde{X}'\widetilde{X})^{-1}\widetilde{X}'\widetilde{y}, \sigma_\varepsilon^2(\widetilde{X}'\widetilde{X})^{-1}). \tag{5.15}$$

Therefore, the posterior for β given $\widetilde{y}$ is known except for σ_ε^2, but this is sufficient for the use as a prior in the Kalman filter with the following measurement and transition equation

$$\overline{y}_{.t} = \begin{pmatrix} 1 & \overline{D}_{.t} & t \times \mathbf{1}_{n_t} & \overline{X}_{.t} \end{pmatrix} \begin{pmatrix} \mu_t \\ \vartheta_t \\ \kappa_t \\ \beta_t \end{pmatrix} + \overline{\varepsilon}_{.t} \tag{5.16}$$

and

$$\begin{pmatrix} \mu_{t+1} \\ \vartheta_{t+1} \\ \kappa_{t+1} \\ \beta_{t+1} \end{pmatrix} = \begin{pmatrix} \mu_t \\ \vartheta_t \\ \kappa_t \\ \beta_t \end{pmatrix} + \begin{pmatrix} \eta_t \\ \omega_t \\ 0 \\ 0 \end{pmatrix}, \tag{5.17}$$

with $\overline{\varepsilon}_{.jt} \sim N(0,\ \sigma_\varepsilon^2/n_{jt})$. The filter is initialized by the informative prior $\beta_0 \sim f(\beta|\widetilde{y}, \sigma_\varepsilon^2)$. This approach reduces the number of datapoints considerably from N to d and avoids the initialization problem of β. Alternatively, a diffuse Kalman filter can be used with only μ_t and ϑ_t in the state vector.

We are also interested in the estimates of Ψ. The Kalman filter for (5.16) and (5.17) produces a likelihood $f(\overline{y}|\widetilde{y}, \Psi)$. This likelihood is

$$\begin{aligned} f(\overline{y}|\widetilde{y}, \Psi) &= \int f(\overline{y}|\mu, \vartheta, \kappa, \beta, [\widetilde{y}], \Psi)f(\mu, \vartheta, \kappa, \beta|\widetilde{y}, \Psi)\mathrm{d}\mu\mathrm{d}\vartheta\mathrm{d}\beta \\ &= \int f(\overline{y}|\mu, \vartheta, \kappa, \beta, \Psi)\pi(\mu|\Psi)\pi(\vartheta|\Psi)\pi(\kappa|\Psi)f(\beta|\widetilde{y}, \sigma_\varepsilon^2)\mathrm{d}\mu\mathrm{d}\vartheta\mathrm{d}\kappa\mathrm{d}\beta. \end{aligned}$$

The basis for inference on Ψ is the posterior $f(\Psi|y)$ that is proportional to the prior $\pi(\Psi)$ times the likelihood $f(y|\Psi)$. The total likelihood $f(y|\Psi)$ is

$$f(y|\Psi) = f(\overline{y}, \widetilde{y}|\Psi) = f(\overline{y}|\widetilde{y}, \Psi) f(\widetilde{y}|\sigma_\varepsilon^2). \tag{5.18}$$

$f(\widetilde{y}|\sigma_\varepsilon^2)$ is the likelihood that results from regressing the equation

$$\widetilde{y} = \widetilde{X}\beta + \widetilde{\varepsilon} \tag{5.19}$$

and is proportional to the distribution of the sufficient statistic $s^2 = \widetilde{y}' M_{\widetilde{X}} \widetilde{y}/(n-d-k)$, with $M_{\widetilde{X}} = (I - \widetilde{X}(\widetilde{X}'\widetilde{X})^{-1}\widetilde{X}')$; see lemma 4.1 of Box and Taio (1973). So $f(\widetilde{y}|\sigma_\varepsilon^2)$ can be written as

$$f(\widetilde{y}|\sigma_\varepsilon^2) \propto {\sigma_\varepsilon}^{-(n-d-k)} \exp(-\frac{1}{2\sigma_\varepsilon^2} \widetilde{y}' M_{\widetilde{X}} \widetilde{y}). \tag{5.20}$$

Usually σ_ε^2 is concentrated out of the likelihood function (5.18) by writing σ_ε^2 as a function of $\{q_1, q_2\}$. In this case this approach is also possible because the prior for β is known up to the scaling factor σ_ε^2.

ML estimators of Ψ are obtained by optimizing the concentrated likelihood with respect to the parameters $\{q_1, q_2\}$. The likelihood can also be used for posterior mode estimation of Ψ with, for example, flat priors for Ψ. In both cases, smoothed estimates of μ_t, ϑ_t, κ, and β are obtained by replacing Ψ by these estimates of Ψ. The derivation of smoothed estimates of μ_t, ϑ_t, κ, and β is presented in Section 5.4.

5.2.5 Summary

So far the procedure can be summarized as follows:

1. Estimate β by means of OLS from the equation in deviations from the means per time per cluster (5.19).

2. Use this estimate of β as a prior for the Kalman filter with equations in means per cluster per time ((5.16) and (5.17)).

3. The total likelihood is simply the product of the likelihood resulting from regression and the likelihood resulting from the Kalman filter (5.18).

This approach has three main advantages. The first is that the number of datapoints in the Kalman filter is reduced considerably. The second is that an initialization problem of β is circumvented. Another advantage is that investigation in the functional form of X can be done very simply by concentrating on the regression equation if most of the information is contained in this equation.

5.2.6 A full Bayesian estimation

A full Bayesian estimation of μ, ϑ, κ, β, and Ψ requires the use of the Gibbs sampler, so we need $f(\alpha|y,\Psi)$ and $f(\Psi|y,\alpha)$ with $\alpha = \begin{pmatrix} \mu' & \vartheta' & \kappa' & \beta' \end{pmatrix}'$. The joint density of α is written as

$$f(\alpha|y,\Psi) = f(\alpha_T|Y_T,\Psi)\prod_{t=1}^{T-1} f(\alpha_t|\alpha_{t+1},Y_t,\Psi).$$

α is simulated from the conditional densities. This is done in the following way. First the two-stage procedure conditional on Ψ is run. From $f(\alpha_T|Y_T,\Psi)$, μ_T, ϑ_T, κ, and β are sampled. Note that the third and fourth component of $\alpha_t|\alpha_{t+1},\Psi$ are deterministic, namely, equal to κ and β. Therefore, only $\mu_t,\vartheta_t|\mu_{t+1},\vartheta_{t+1},Y_t,\Psi$ needs to be sampled from a normal density $N(x^*_{t|T},P^*_{t|T})$ for $t = T-1,\ldots,1$. The moments are given by proposition 2 of Frühwirth-Schnatter (1994) and are

$$\begin{aligned}
x^*_{t|T} &= \begin{pmatrix} \mathrm{E}[\mu_t|Y_t] \\ \mathrm{E}[\vartheta_t|Y_t] \end{pmatrix} + A_{t+1}\begin{pmatrix} \mu_{t+1} - \mathrm{E}[\mu_t|Y_t] \\ \vartheta_{t+1} - \mathrm{E}[\vartheta_t|Y_t] \end{pmatrix}, \\
P^*_{t|T} &= \sigma^2_\varepsilon(I_{B+1} - A_{t+1})\Sigma_{t+1}, \\
\Sigma_{t+1} &= \begin{pmatrix} \mathrm{Var}(\mu_t|Y_t) & \mathrm{Cov}(\vartheta_t,\mu_t|Y_t) \\ \mathrm{Cov}(\mu_t,\vartheta_t|Y_t) & \mathrm{Var}(\vartheta_t|Y_t) \end{pmatrix},
\end{aligned}$$

and

$$A_{t+1} = \Sigma_{t+1}\left\{\Sigma_{t+1} + \begin{pmatrix} q_1 & 0 \\ 0 & q_2 I_B \end{pmatrix}\right\}^{-1}.$$

The density $f(\Psi|y,\alpha)$ is derived in the following way:

$$f(\Psi|y,\alpha) = \prod_{t=1}^{T} f(y_t|\alpha_t,\Psi)\prod_{t=1}^{T} f(\alpha_t|\alpha_{t-1},\Psi)f(\Psi).$$

For more details of Gibbs sampling, see, for example, Carter and Kohn (1994), Frühwirth-Schnatter (1994), and Shephard (1994).

5.3 The initialization of cluster trends

So far no attention has been paid to the initialization of ϑ. It is possible that the first observation in a cluster is nearly at the end of the period, so parts of the state vector are for a long time undetermined. In our example it is not possible to rely on a "large k" approximation because of numerical problems. It is possible to use the diffuse Kalman filter of De Jong (1991a), but in this case a more simple approach is possible by sacrificing the first observation in a cluster

to initialize the cluster trend (see e.g., Harvey (1989, ex. 3.2.1)).

The state vector at time t can be divided into three parts, a part where observations are available at time $t-1$ (A), a part where at time t the first observation is available (F), and finally a part where at time t no observations are available (N). In the filter, part N of the state can be ignored in the prediction and update steps as well as in the likelihood evaluation. If $j \in F$, the state vector is extended. The observations are not used to evaluate the likelihood; actually the likelihood computed is conditional on these observations (a standard procedure in the Kalman filter).

Because $\vartheta_{jt} = \overline{y}_{.jt} - \overline{x}^*_{.jt}\beta^*_t - \overline{\varepsilon}_{.jt}$, the expectation and covariance matrix of the extended state follow from

$$\begin{aligned}
\mathrm{E}\left[\vartheta_{jt}|\overline{Y}_{.t}\right] &= \overline{y}_{.jt} - \overline{x}^*_{.jt}\,\mathrm{E}\left[\beta^*_t|\overline{Y}_{.t}\right], \\
\mathrm{Var}\left(\vartheta_{jt}|\overline{Y}_{.t}\right) &= \overline{x}^*_{.jt}\,\mathrm{Var}\left(\beta^*_t|\overline{Y}_{.t}\right)\overline{x}^{*\prime}_{.jt} + \sigma^2_\varepsilon/n_{jt}, \\
\mathrm{Cov}(\vartheta_{jt},\vartheta_{mt}|\overline{Y}_{.t}) &= \overline{x}^*_{.jt}\,\mathrm{Var}\left(\beta^*_t|\overline{Y}_{.t}\right)\overline{x}^{*\prime}_{.mt} \text{ if } m \in F, \\
&= -\overline{x}^*_{.jt}\,\mathrm{Cov}(\beta^*_t,\vartheta_{mt}|\overline{Y}_{.t}) \text{ if } m \in A,
\end{aligned}$$

and

$$\mathrm{Cov}(\vartheta_{jt},\beta^*_t|\overline{Y}_{.t}) = -\overline{x}^*_{.jt}\,\mathrm{Var}(\beta^*_t|\overline{Y}_{.t}),$$

with $j \in F$, $\overline{x}^*_{.jt} = \begin{pmatrix} 1 & t & \overline{x}'_{.jt} \end{pmatrix}'$, $\beta^*_t = \begin{pmatrix} \mu_t & \kappa & \beta'_t \end{pmatrix}'$, $\overline{Y}_{.t} = \{\overline{y}_{.1},\ldots,\overline{y}_{.t}\}$, and conditional on Ψ. Together with the updated part of the state A, these formulas provide the result required for further recursion.

5.4 Smoothing with update and downdate

Smoothed estimators can be obtained by the fixed-interval classical smoothing algorithms. A not very well-known smoothing algorithm, the forward-backward algorithm shown by Merkus, Pollock, and de Vos (1993) and Kitagawa (1994), is used in this chapter. This filter is another example of the use of Bayesian technology in the Kalman filter. It uses the estimates of $a_{t|t-1}$ and $P_{t|t-1}$ and the estimates $\widetilde{a}_t$ and $\widetilde{P}_t$ of the inverse filter with $a_{t|s} = \mathrm{E}[\alpha_t|y_1,\ldots,y_s]$, $P_{t|s} = \mathrm{E}[(\alpha_t - a_s)(\alpha_t - a_s)']$, $s \le t$, and $\widetilde{a}_{t|s} = \mathrm{E}[\alpha_t|y_s,\ldots,y_T]$, $\widetilde{P}_{t|s} = \mathrm{E}[(\alpha_t - \widetilde{a}_s)(\alpha_t - \widetilde{a}_s)']$, $s \ge t$. The inverse filter uses the same measurement equation as the normal Kalman filter. The transition equation is the inverse of the normal transition equation. The data are used in backward order. If prior information is noninformative, then the smoothed estimate of $a_{t|T}$ is a weighted average of $a_{t|t-1}$ and $\widetilde{a}_t$ weighted with $P_{t|t-1}$ and $\widetilde{P}_t$:

$$\begin{aligned}
P_{t|T} &= (P^{-1}_{t|t-1} + \widetilde{P}^{-1}_t)^{-1}, \\
a_{t|T} &= P_{t|T}(P^{-1}_{t|t-1}a_{t|t-1} + \widetilde{P}^{-1}_t\widetilde{a}_t).
\end{aligned}$$

This may not be the most efficient algorithm, but it is intuitively clear and rather straightforward to program. It combines all the information "from the left" and "from the right." It is also clear how to proceed in cases in which parts of the state vector and the covariance matrix are not defined.

A slight complication arises in the general case, in which the state vector is α_t and the prior information for α_t, say $\alpha_0 \sim N(a_0, P_0)$, is used for both the update and the downdate. Then the prior at time t has expectation $a_{t|0}$ and variance $P_{t|0}$. The update and downdate estimators $a_{t|t-1}$, $P_{t|t-1}$, $\widetilde{a}_t$, and $\widetilde{P}_t$ include the prior information. It can be demonstrated that the prior information must be subtracted once (De Vos and Merkus (1992)), so

$$P_{t|T} = (P_{t|t-1}^{-1} + \widetilde{P}_t^{-1} - P_{t|0}^{-1})^{-1} \tag{5.21}$$

and

$$a_{t|T} = P_{t|T}(P_{t|t-1}^{-1} a_{t|t-1} + \widetilde{P}_t^{-1}\widetilde{a}_t - P_{t|0}^{-1} a_{t|0}). \tag{5.22}$$

Consider the model of the last section with the state variables $\mu_t, \vartheta_t, \kappa_t$, and β_t. For μ_t, κ_t, and β_t, prior information is used. The prior information for β is the same at any time, $\beta_t \sim N((\widetilde{X}'\widetilde{X})^{-1}\widetilde{X}'\widetilde{y}, \sigma_\varepsilon^2(\widetilde{X}'\widetilde{X})^{-1})$; see (5.15). The prior information for κ_t is $\kappa_t \sim N(\kappa_0, \sigma_\varepsilon^2 K_0)$. Prior information for μ_t differs in time, $\mu_t|\text{prior} \sim N(0, (t-1)\sigma_\varepsilon^2 q_1)$. For ϑ_t, no prior information is used. So the prior information can be written as

$$a_{t|0} = \begin{pmatrix} 0 \\ 0 \\ \kappa_0 \\ (\widetilde{X}'\widetilde{X})^{-1}\widetilde{X}'\widetilde{y} \end{pmatrix}, \qquad P_{t|0} = \sigma_\varepsilon^2 \begin{pmatrix} (t-1)q_1 & 0 & 0 & 0 \\ 0 & C & 0 & 0 \\ 0 & 0 & K_0 & 0 \\ 0 & 0 & 0 & (\widetilde{X}'\widetilde{X})^{-1} \end{pmatrix},$$

with $C^{-1} \to 0$.

From the $\text{Cov}(\alpha_t, \alpha_{t+1}|\text{prior})$ with $\alpha_t = \begin{pmatrix} \mu_t & \vartheta_t' & \kappa_t & \beta_t' \end{pmatrix}'$, the inverse transition equation is calculated. It follows from the conditional bivariate normal density that

$$\text{E}[\alpha_t|\alpha_{t+1}, \text{prior}] = \begin{pmatrix} \frac{t-1}{t} & 0 & 0 & 0 \\ 0 & I_B & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & I_k \end{pmatrix},$$

$$\text{Var}(\alpha_t|\alpha_{t+1}, \text{prior}) = \sigma_\varepsilon^2 \begin{pmatrix} \frac{t-1}{t}q_1 & 0 & 0 & 0 \\ 0 & q_2 I_B & 0 & 0 \\ 0 & 0 & K_0 & 0 \\ 0 & 0 & 0 & (\widetilde{X}'\widetilde{X})^{-1} \end{pmatrix}.$$

Estimates of $\widetilde{a}_t$ and $\widetilde{P}_t$ are obtained by this inverse transition equation and the measurement equation. The smoothed estimates follow from (5.22) and (5.21).

5.5 The application: housing prices in Amsterdam

In this section an application of a model for stochastic hierarchical trends is applied to the example of apartment selling prices. Selling prices for the existing stock of apartments are compiled from a database from Gemeente Amsterdam Dienst Belastingen (the department of municipal taxes, Amsterdam). The sales were realized in the period from January 1986 up to and including August 1996. They have been screened extensively. For example, sales between relatives are omitted. In 76 out of 93 neighborhoods in Amsterdam, 12,716 sales have been realized.

The database also contains the characteristics of the sold apartments, such as age, surface of the living area, surface of the garden, presence of an elevator in the building, marks for the quality and maintenance situation, and the fact the apartment is situated on a canal.

The development of selling prices follows some general pattern, but housing price trends may vary over different neighborhoods. For this reason a distinction is made between a general trend and trends per neighborhood. The general trend is denoted by the variable μ_t and concerns the general increase or decrease in selling prices in time. The variable ϑ_{jt} is the deviation from the general trend for neighborhood j. The sum of μ_t and ϑ_{jt} provides the trend for neighborhood j. The explanatory variables x_{ijt} account for differences in individual characteristics of the houses.

The model variables are shown in Table 5.1. A distinction is made between old apartments in the center and outside the center because old apartments in the centre are, in many cases, listed buildings. The explanatory variables are specified as $\beta_1 \ln(\text{Living}+\beta_2\text{Garden}) + \beta_3\text{A1900C} + \beta_4\text{A1900N} + \beta_5\text{Age20} + \beta_6\text{Age45} + \beta_7\text{Age} + \beta_8\text{Elevator} + \beta_9\text{Canal} + \beta_{10}\ln(\text{Quality}) + \beta_{11}\ln(\text{Maintain})$: $\ln(\text{Living}+\beta_2\text{Garden})$ can be linearized by choosing a good starting value for β_2 and realizing that $\ln(1+\varepsilon) \simeq \varepsilon$ if ε is small. The dependent variable is the logarithm of the selling price. The model specification is the same as (5.6) to (5.8). A prior distribution is used for κ_t, $\kappa_t \sim N(0, .1)$.

First, the model is estimated by regression on the observations in deviations from the means per time and neighborhood. In this stage, after many experiments the functional form is chosen. Results from this regression are shown in Table 5.2. The ultimate results, which are nearly the same as those from regression, are shown in Table 5.3. Apparently there is not much information in the means. Note that in this way a check is provided whether the specification search in the first stage may be supposed to have led to the optimal result. This is an example of the applicability of our algorithm in complex models.

Table 5.1: Model variables.

Variable	Description
Sale	Selling price
Living	Square meters of living area
Garden	Square meters of garden
A1900C	$\begin{cases} 1 & \text{if year of building} < 1900 \text{ an, apartment is situated in the center} \\ 0 & \text{else} \end{cases}$
A1900N	$\begin{cases} 1 & \text{if year of building} < 1900 \text{ an, apartment is situated outside the center} \\ 0 & \text{else} \end{cases}$
AGE20	$\begin{cases} 1 & \text{if } 1900 \leq \text{ year of building} < 1920 \\ 0 & \text{else} \end{cases}$
AGE45	$\begin{cases} 1 & \text{if } 1920 \leq \text{ year of building} < 1945 \\ 0 & \text{else} \end{cases}$
AGE	$\begin{cases} 1 & \text{year of sale – year of building, if year of building} \geq 1945 \\ 0 & \text{else} \end{cases}$
Elevator	$\begin{cases} 1 & \text{if an elevator is present in the building} \\ 0 & \text{else} \end{cases}$
Canal	$\begin{cases} 1 & \text{if apartment is situated at a canal} \\ 0 & \text{else} \end{cases}$
Quality	An indicator for quality $\{3, 4, \ldots, 10\}$
Maintain	An indicator for maintainance $\{3, 4, \ldots, 10\}$
μ_t	The general trend at time t
ϑ_{jt}	The trend in neibhorhood j at time t
time	Time in months, $t = 0$ equals January 1, 1986

Table 5.2: Results from Regression.

Variable	Coefficient	Standard error	T value
Living	.8172	.0065	125.13
Garden	.1220	.0132	9.22
A1900C	-.1269	.0119	-10.65
A1900N	-.1507	.0150	-10.03
Age20	-.1202	.0152	-7.90
Age45	-.0992	.0143	-6.91
Age	-.0041	.0004	-9.62
Elevator	.0361	.0076	4.77
Canal	.1061	.0072	14.77
Quality	.4226	.0278	15.20
Maintain	.2409	.0242	9.98
no.	8185		
$\widehat{\sigma}_{\varepsilon}$	.1757		

Table 5.3: Results from the Kalman filter.

Variable	Coefficient	Standard error	T value
Living	.8188	.0054	152.70
Garden	.1294	.0105	12.38
A1900C	-.1178	.0096	-12.29
A1900N	-.1407	.0112	-12.53
Age20	-.1244	.0116	-10.76
Age45	-.0991	.0112	-8.86
Age	-.0037	.0003	-11.61
Elevator	.0414	.0062	6.69
Canal	.1144	.0058	19.56
Quality	.4135	.0219	18.90
Maintain	.2381	.0190	12.56
Time	.0069	.0006	10.78
no.	12716		
$\widehat{\sigma}_{\varepsilon}$	.1825		
$\widehat{q}_1$	.0012		
$\widehat{q}_2$	.0094		

Figure 5.1 gives the general trend and figure 5.2 an example of the trend for a specific neighborhood as a deviation from the general trend. The y axis is in logarithms, so an increase of .1 means that the increase in selling prices is approximately 10%.

The dashed lines indicate the 95% confidence intervals and the points the standardized selling prices, corrected for individual characteristics and the general trend, $\overline{y}_{.jt} - \mu_t - t\widehat{\kappa} - \overline{x}_{.jt}\widehat{\beta}$. Figure 5.2 shows that for the specific neighborhood the selling prices increase more than the general trend.

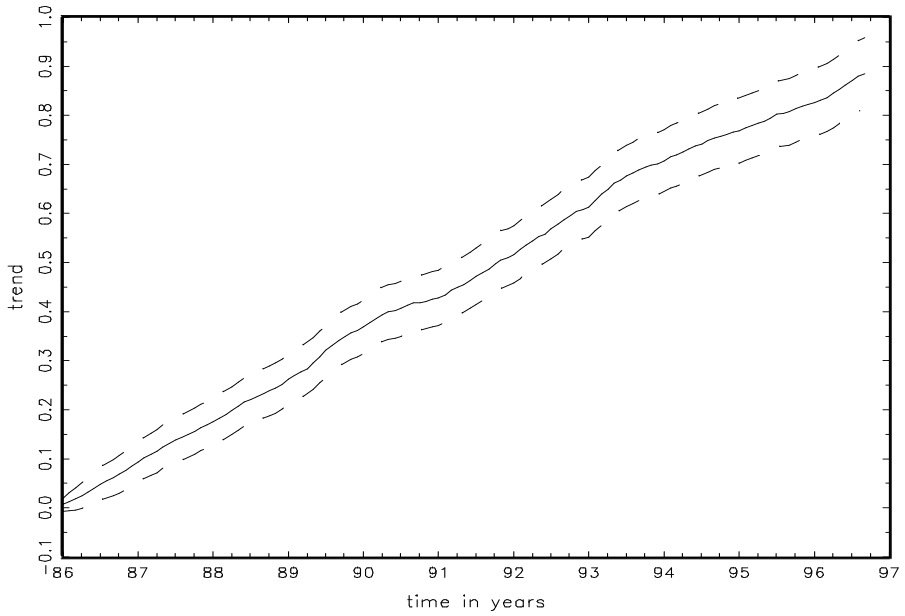

Figure 5.1: General trend.

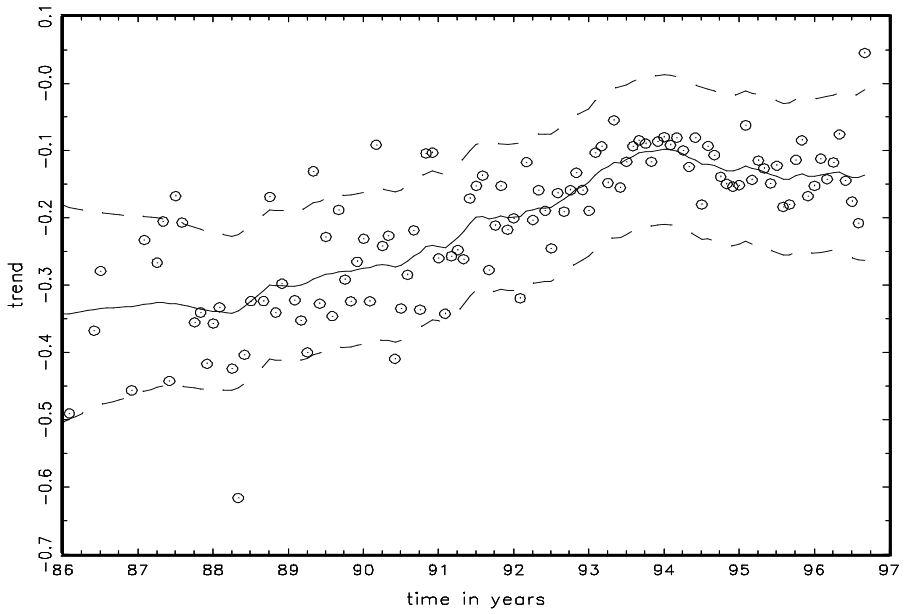

Figure 5.2: Trend for a specific neighborhood.

Chapter 6

The hierarchical trend model

Abstract[1]

This chapter presents a Hierarchical Trend Model for selling prices of houses, addressing three main problems: the spatial and temporal dependence of selling prices and the dependency of price index changes on housing quality. In this model the general price trend, cluster-level price trends, and specific characteristics play a role. Every cluster, a combination of district and house type, has its own price development. The Hierarchical Trend Model is used for property valuation and for determining local price indices. Two applications are provided, one for the Breda region, and one for the Amsterdam region, lying respectively south and north in the Netherlands. For houses in these regions the accuracy of the valuation results are presented together with the price index results. Price indices based on the Hierarchical Trend Model are compared to a standard hedonic index and an index based on weighted median selling prices published by national brokerage organization. It is shown that, especially for small housing market segments the Hierarchical Trend Model produces price indices which are more accurate, detailed, and up-to-date.

6.1 Introduction

This chapter concerns the modeling of selling prices of houses by hedonic price models. Besides the size and the location of a house, the selling date is an important characteristic to explain selling prices in a time of rapid price movements. A Hierarchical Trend Model (HTM) is presented, addressing the spatial and the temporal dependence of selling prices.

In literature on hedonic modeling the temporal dependence of selling prices is addressed in several ways. Case and Quigley (1991) present a model in which information on repeated sales is combined with that of single sales. Discrete and continuous time varying locational and structural parameters are considered. In the discrete case for every time period an extra parameter is added to the model. In the continuous case, the model specification is more parsimonious, but a linear trend for parameter evolution is imposed. Fleming and Nellis (1992)

[1]This chapter is based on Francke and Vos (2004).

propose repeated regressions for every time period, so regression parameters may vary over time. A closely related approach is provided by Knight, Dombrow, and Sirmans (1995). In this approach only one regression is performed with varying parameters over time, and correlation between multiple sales is considered.

As indicated by Schwann (1998) the problem of these methods is the lack of degrees of freedom, because an extra parameter is added to the model for every time period per characteristic. In the case of thin markets a small number of observations may lead to unreliable parameter estimates. The weakness in these methods is that it is assumed that parameter values in one period do not affect parameter values in other periods. For that reason Schwann proposes a time series model in which a stochastic structure for the parameter evolution is assumed. In his model the only time varying variable is the constant, but there is no reason to restrict it to this variable. It can be applied to locational and structural variables as well.

The HTM can be seen as an extension of the time series model proposed by Schwann. In the HTM parameters vary over time, location, and house type. In the HTM, which can be described as a dynamic hedonic price model, cluster-level price trends, a general price trend, and specific characteristics play a role. Together with the influence of the specific characteristics these trends are estimated within the HTM. The cluster-level trends are specified as deviations from the general trend. The general and cluster-level trends are modeled as stochastic trends, for example by random walks. The clusters, or market segments can be defined by (a combination of) districts and house types. In this set-up it is possible that every market segment has a different price development. A closely related approach is provided by Gelfland et al. (1998). This study concentrates on spatio-temporal modeling of residential sales data in a Bayesian framework, but no house type trends are considered.

Sections $6.2 - 6.4$, and 6.8 concern model specification. In section 6.2 the functional form of the dependent variable is motivated. In section 6.3 the choice of the functional form of some of the explanatory variables, like lot and house size, is discussed. Section 6.4 describes the Hierarchical Trend Model. Section 6.8 provides some temporal and spatial model modifications and estimation results.

A first application of this model is found in the valuation of property, in this case individual houses. Given the characteristics of a house the model is able to produce values for all time points in the time period considered. At this moment the HTM is operational in Amsterdam and several other Dutch cities for the determination of local real estate taxes.

A second application is found in the determination of price indices. Changes in the levels of selling prices can be caused by changes in the underlying characteristics of sold houses. For this reason selling price levels in one period cannot be compared directly to selling price levels in another period, but the levels must be adjusted for differences in house characteristics. It will be shown that for thin markets the estimated price trends from the HTM provide the correct measurement of house price movements, and levels over time.

In this chapter the HTM is estimated for two datasets over the period $1985 - 1999$, both

from the Dutch Broker Organization (NVM). The first dataset contains selling prices for the Breda region, the second database contains selling prices for the Amsterdam region. Section 6.5 provides a brief description of both datasets.

Section 6.6 presents specific valuation results for the Breda region. In section 6.7 price indices are shown for both the Breda and the Amsterdam region. The indices produced by the HTM are compared to indices obtained from a standard hedonic method as well as from an often-used method that simply consists of taking averages, or medians for every cluster. The latter method is published in reports on price development by the Dutch Brokerage Organization NVM. For all methods standard deviations are compared for price indices for the region as a whole as well as for small market segments within the region, both on a monthly, quarterly, and a yearly basis. Section 6.9 concludes with the key results.

6.2 Dependent variable

In this section the specification of the dependent variable is motivated. The dependent variable is the selling price, or a transformation of the selling price. Examples of transformations are the square root, and the natural logarithm of the selling price. The Box-Cox method is often used as a guideline to choose a specific transformation, see for example Halvorsen and Pollakowski (1981). Let Y_i denote the selling price of sale i for $i = 1, \ldots, n$. The Box-Cox transformation is given by $Y_i(\theta) = ((Y_i)^\theta - 1)/\theta$. For $\theta = 1$, the dependent variable is the selling price, for $\theta = \frac{1}{2}$, the dependent variable is the square root of the selling price, and for $\theta \to 0$, the dependent variable is the natural logarithm of the selling price. In general θ is unknown, and along with the coefficients of the explanatory variables, it needs to be estimated.

We did not use the Box-Cox analysis to choose a transformation. We use the natural logarithm of the selling price as dependent variable. The reason for this is that we assume that variables for districts and trends work in a multiplicative way on the size of the house, see section 6.3. Another reason is that our goal is to minimize the relative standard deviation, see appendix A.6. This can also be done in the more general cases of a Box-Cox transformation, but is more complex to evaluate. An additional assumption of the natural logarithm is that the error terms have a lognormal density, which can be checked by evaluating the residuals.

6.3 Multiplicative/additive model

As shown by Halvorsen and Pollakowski (1981) the appropriate functional form cannot in general be specified on theoretical grounds, but it is a matter of convenience. Several functional forms are used in practice: linear, semi-log, log-linear and inverse semi-log, see Palmquist (1984). We use the log-linear model specified as

$$y = \beta_1 \ln x_1 + Z\delta + \varepsilon,$$

with y the natural log of the selling price, x_1 the internal floorspace, and Z being the other explanatory variables. In this specification an increase of x_1 by 1 percent will result in an increase of Y of approximately β_1 percent. It is expected that $\beta_1 < 1$, so the value will be less than proportional with the internal floorspace.

Another important characteristic is the lot size (x_2). If the natural log of the lot size is added as independent variable, then

$$Y = {x_1}^{\beta_1} {x_2}^{\beta_2} \exp(Z\delta + \varepsilon), \tag{6.1}$$

with Y the selling price. So, in this example (a power of) the internal floorspace is multiplied by (a power of) the lot size. This feature of the model is regarded by real estate agents and valuers as undesirable: the mutual influence of floorspace and lot size is expected to be additive, rather than multiplicative. For that reason we change the model specification in

$$y = \alpha \ln(X\beta) + Z\delta + \varepsilon, \tag{6.2}$$

where β denotes a $k \times 1$ vector of coefficients of additive variables $X = \begin{bmatrix} x_1 & \cdots & x_k \end{bmatrix}$, and Z is a matrix of other explanatory variables [2]. Note that the coefficient β_1 for x_1 is 1, because otherwise where a constant is included in Z, the model is not identified. If we take the exponent for this model, we get

$$Y = (X\beta)^{\alpha} \exp(Z\delta + \varepsilon), \tag{6.3}$$

so this model is additive in X.

The variables Z as a factor influence both the lot size and internal floorspace, as is apparent from (6.3). This is a desirable feature for variables concerning time trends, and the influence of the district. A disadvantage of the specification (6.2) is that variables like the age of the building and the maintenance both influence the value of the floorspace, and the lot size, instead of just the floorspace, but we will not pursue this issue here.

Because equation (6.2) is nonlinear it cannot be estimated by Ordinary Least Squares (OLS). In appendix A.7 an estimation procedure is provided.

[2] In practice estimation results show no difference in goodness of fit for both specifications (6.1) and (6.2).

It is possible to linearize (6.2) in X, so the model value M_j can be written as $M_j = \sum_{i=1}^{k} x_{ij}\phi_{ij}$. From (6.2) it follows that

$$M_j = c\left(\sum_{i=1}^{k} x_{ij}\beta_i\right)^{\alpha-1}\left(\sum_{i=1}^{k} x_{ij}\beta_i\right) = \sum_{i=1}^{k}\left[\frac{c\beta_i}{\left(\sum_{i=1}^{k} x_{ij}\beta_i\right)^{1-\alpha}}\right] x_{ij} = \sum_{i=1}^{k} x_{ij}\phi_{ij},$$

$$\phi_{ij} = \frac{\beta_i}{\sum_{i=1}^{k} x_{ij}\beta_i} M_j,$$

with $c = \exp(z_j\delta)$, with z_j row j of Z.

6.4 The hierarchical trend model

6.4.1 Hierarchical trends

The Hierarchical Trend Model is a dynamic model for selling prices of houses. In this model individual characteristics and price trends play a role. Cluster-level price trends are distinguished from a general price trend. Examples of clusters are districts and house types. The general trend, and the cluster-level as deviations from the general trend, are modeled as stochastic trends.

Let the vector y_t represent the logarithms of the selling prices of houses at time t. We denote the length of y_t by n_t, and the k-th observation in y_t by y_{kt} $(k = 1, \ldots, n_t)$. First, we assume that all prices follow a common trend, the general trend, which we can write as

$$y_t = \mathbf{i}\mu_t + \epsilon_t, \tag{6.4}$$

where $\mathbf{i}$ is a n_t-vector of ones, and $\epsilon_t \sim N(0, \sigma^2 I)$, with I a $n_t \times n_t$ identity matrix. Note that we have suppressed the time dependency of $\mathbf{i}$ and I in the notation. μ_t is a scalar stochastic trend process.

The general trend can be specified in several ways. An example of a stochastic process is the random walk with drift,

$$\mu_{t+1} = \mu_t + \kappa + \eta_t, \qquad \eta_t \sim N(0, q_1\sigma^2), \tag{6.5}$$

with a given μ_1. The disturbances ϵ and η are assumed to be independent. Note that in case explanatory variables are added, for $q_1\sigma^2 \to \infty$, equations (6.4) and (6.5) specialize to a fixed effects model. If we let $q_1\sigma^2 \to 0$ this reduces to a straight line with slope κ.

Suppose we have a method to categorize houses into L different types. We can include a dummy matrix D_t for house types as regressors in the model. Each row in the $n_t \times L$ matrix D_t has a one in the l-th column and zeros elsewhere, if the house is of type $l = 1, \ldots, L$. Writing

λ for the regression parameter vector for house types,

$$y_t = \mathbf{i}\mu_t + D_t\lambda + \epsilon_t. \tag{6.6}$$

In this model, the relative price differences between house types stay constant through time. If we expect the prices to grow at different rates, we could allow λ to vary over time. The elements of the vector λ_t can be modeled as trends in a similar fashion as the common trend μ_t. The specifications for the house type trends are typically less elaborate than for the common trend; we will model them as simple independent random walks, with a common variance level.

We can see immediately that if both λ and μ are constant, an unrestricted specification like (6.6) leads to the dummy trap. In the general time-varying case, there is also an identification problem if we try to extract both a general trend and a trend for house types from the data. We can solve this by imposing the restriction $\mu_1 = 0$. With this restriction, the level of μ_t indicates the general price increase relative to the first time period. A trend for a specific house type is obtained as the sum of μ_t and the element of λ_t of corresponding to the house type.[3]

Of course, there is no need to restrict this approach to house types; any qualitative independent variable can be treated in this way. We will refer to these as clusters. An obvious example is a variable which indicates the district where the house is located. Note that if we model two classifications simultaneously (e.g. house types and districts), additional restrictions are required.

We model the vector of log house prices with an extended version of (6.6), the HTM:

$$y_t = \mathbf{i}\mu_t + D_t\lambda_t + \dot{D}_t\vartheta_t + \ddot{D}_t\phi + X_t\beta + \epsilon_t, \qquad \epsilon_t \sim N(0, \sigma^2 I). \tag{6.7}$$

We specify the general trend μ_t as a random walk with drift (6.5). The vector λ_t contains trend levels for house types at time t, while the vector ϑ_t contains trend levels for districts. The matrices D and $\dot{D}$ contain ones and zeros such that they select the appropriate house type and district for the observation. For now, we assume random walks for these trends:

$$\lambda_{t+1} = \lambda_t + \varsigma_t, \qquad \varsigma_t \sim N(0, q_2\sigma^2 I), \tag{6.8}$$

$$\vartheta_{t+1} = \vartheta_t + \omega_t, \qquad \omega_t \sim N(0, q_3\sigma^2 I), \tag{6.9}$$

where the identity matrices I have the appropriate dimensions.

Each district is divided in a number of neighborhoods, for which we assume separate levels. We collect all levels in a vector ϕ, and use a selection matrix $\ddot{D}$ to assign the appropriate neighborhood level to the observations. We can treat the levels as fixed or random effects. We

[3] An alternative solution is to drop the common trend from the model. Without a common trend, the correlations between the house type trends will have to be specified through the disturbance variance matrix.

propose a random effect specification,

$$\phi \sim N(0, q_4\sigma^2 I). \tag{6.10}$$

Finally, we add a number of explanatory variables X_t with fixed parameters. We will keep the basic form of the model linear, so a specification like $(\ln x_t'\beta)^\alpha$ will be approximated by an iterative procedure as described in appendix A.7.

The complete model specification is provided by the equations (6.5), and (6.7)–(6.10). Note that in (6.7) β is kept constant over time and over clusters, and homoscedasticity is assumed. In equations (6.8), and (6.9) the variances are also kept constant over the clusters. These assumptions will be relaxed in section 6.8.

6.4.2 Structural time series model

In the method of time series modeling we described, observations are assumed to be aggregates of unobserved parts with some interpretation, such as trend, and cycle. Each part can be modeled further with as much detail as desired. These models are known in the literature as *Structural time series*, or *Unobserved components* models. For a detailed description we refer to Harvey (1989), West and Harrisson (1997), and Durbin and Koopman (2001), who discuss these models as examples of *state-space* or *Dynamic linear* models. In the state-space form, the unobserved components can be estimated with the *Kalman filter* algorithm.

A model in state-space format consists of a measurement and a transition equation. The measurement equation relates the unknown state vector α_t to the observations y_t. The transition equation describes the evolution in time of the state vector α_t.

To put the model into state-space format, we stack the variables μ_t, κ, and the vectors $\lambda_t, \vartheta_t, \phi, \beta$ in the state vector α_t. The measurement equation is simply

$$y_t = Z_t\alpha_t + \epsilon_t = [\ \mathbf{i}\quad 0\quad D\quad \dot{D}\quad \ddot{D}\quad X_t\ ]\alpha_t + \epsilon_t, \qquad \epsilon_t \sim N(0, \sigma^2 I). \tag{6.11}$$

In the transition equation

$$\alpha_{t+1} = T_t\alpha_t + \xi_t, \tag{6.12}$$

the transition matrix T_t is a time independent block diagonal matrix, with $\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$ on the upper block, and I on the lower block. The zero-mean Normal transition disturbance ξ_t has a diagonal variance matrix, with

$$\sigma^2[\ q_1\quad 0\quad q_2 \ldots q_2\quad q_3 \ldots q_3\quad 0 \ldots 0\quad 0 \ldots 0\ ]$$

on the diagonal.

6.4.3 Estimation issues

We already mentioned an identification problem in specifying trends on different levels. In the general model (6.7), we will set the initial general trend level μ_0 at zero.

Another identification issue results from the fact that we have two complete classifications for the houses: districts and house types. This can be solved by setting the initial level of some house type at zero. The general trend μ_t is interpreted as general with regard to districts for houses of this type. The vector $\mathbf{i}\mu_t + D_t\lambda_t$ provides the trends for all house types, generally with regard to districts.

Models in state-space format can be estimated by the Kalman filter. The Kalman filter is usually applied to univariate time series models, but it can also be applied to multivariate data, even for situations with an unequal number of observations over time, as is the case in the HTM. The state space formulation of the HTM specified by (6.5), and (6.7)–(6.10) is provided by equations (6.11), and (6.12). If the initial state is known, standard Kalman filter (and smoothing) recursions can be applied to this model, providing estimates of the state vector α_t, and a likelihood. This likelihood is optimized with respect to the unknown variance parameters σ^2, and $q_1, \ldots, q_4$.

The Kalman filter assumes the first and second moment of the initial state α_0 to be known. In general this is not true, because (a part of) the initial state is not known, and therefore diffuse. This leads to an initialization problem which can be solved by the diffuse Kalman filter, or the exact initial Kalman filter. The recursions for the diffuse Kalman filter are provided by De Jong (1991a), the recursions for the exact initial filter are provided by Koopman (1992) and Koopman (1997). In the HTM, nonstationarity of the transition equation, and the presence of explanatory variables lead to a diffuse initial state. For that reason the diffuse Kalman filter of de Jong is applied, a method which is also used by Schwann (1998).

In chapter 5, it is shown how a hierarchical trend model with explanatory variables can be computed efficiently. First, we calculate the means per cluster $\bar{y}_1, \ldots, \bar{y}_T$, and the deviations from these means $\tilde{y}_1, \ldots, \tilde{y}_T$. The length of vector $\bar{y}_t$ is the number of different clusters for which we have observations at time t, while $\tilde{y}_t$ has the same dimension as y_t. Likewise, we calculate means and deviations from means for the explanatory variables. The coefficients of the explanatory variables are time- and cluster invariant, and can be computed by applying OLS on the stacked deviation from mean vectors and matrices $\tilde{y} = [\ \tilde{y}_1' \ \ \ldots \ \ \tilde{y}_T'\ ]'$ and $\tilde{X} = [\ \tilde{X}_1' \ \ \ldots \ \ \tilde{X}_T'\ ]'$. Subsequently, the Kalman filter is ran with the mean data $\bar{y}_t, \bar{X}_t$, and with the OLS estimates as initial mean and variance of the explanatory variables in the state. The likelihood is obtained as the product of the OLS likelihood and the Kalman filter likelihood. This approach reduces the number of observations in the diffuse Kalman filter considerably.

All estimation procedures are written in GAUSS (Aptech Systems, Inc). For likelihood optimization the maximum likelihood library from GAUSS is used.

6.5 Applications

6.5.1 Data description

The HTM-model is applied to two different datasets. The first dataset contains selling prices for houses in the Amsterdam region, an urban district with a relatively high proportion of apartments. The second database contains selling prices of the Breda region, a rural district with one middle-sized city Breda of about 160,000 inhabitants. The Breda region has a relatively high proportion of single-family houses.

The two databases were established by the Dutch Brokerage Organization (NVM). They have several merits from the point of view of this study. First, the size of the database is large, because the number of transactions registered by the NVM is on average more than 60% of all transactions registered by the Land Registry (Kadaster). Sample sizes of these magnitudes undoubtedly provide an adequate foundation for measuring house price changes at a regional level. Secondly, the data concerning house characteristics are more extensive than anything available in this area and this again helps to improve the reliability of the statistical analysis. The available information about house characteristics is summarized below:

1. Purchased price: date of selling, asking price, condition on sales
2. Location: address (street, number, postal code)
3. Housing characteristics:

(a) House type: detached, semi-detached, attached, apartment (with sub-classification)
(b) Tenure: freehold, land leasehold condition
(c) Garage: type of garage
(d) Heating type
(e) House size: area in m3
(f) Plot size: total size in m2
(g) Garden: length and position of garden
(h) Space: number of rooms, kitchen, bathroom, type of living room
(i) Age: year of construction
(j) Physical condition: interior maintenance, exterior maintenance
(k) Marketing period
(l) Listed building

As indicated above, the data refer to transactions at the selling agreement stage as opposed to the notarial act stage. This means that the price information is more up-to-date as an indicator of price movements because of the time lags that occur between the price negotiation stage and the ultimate completion of the transaction at the notarial act - a lag that may extend over several months.

In the next two subsections both databases will be described.

Table 6.1: Number of relevant transactions per house type.

Region Type	Amsterdam Number	Breda Number
Attached	5613	7275
Semi-detached	2562	8591
Detached	595	3460
Apartment	22678	1849
Total	31448	21175

6.5.2 Amsterdam region

A special database was designed to accommodate the various measurement problems associated with house prices. The database covers 44,780 purchase transactions of existing dwellings in the Amsterdam region from January 1985 until July 1999. This market area is composed of four municipalities: Amsterdam, Amstelveen, Diemen and Ouder-Amstel.

Transactions without proper postal code were excluded from the database. For the segmentation in Amstelveen the year of construction is one of the crucial criteria. These restrictions concerning house type, postal code and year of construction results in 42616 usable transactions from the original 44780. To be able to correct for differences in quality of the houses it is necessary to take into account housing characteristics like house size, lot size, year of construction, etc. This leads to extra demands on the data resulting in 31448 usable transactions. Especially in the earlier years less than half of the database could be used.

The sales volume per year doubled during the period $1985-1999Q2$. The fraction of sales in the municipality of Amsterdam alone is more than 75% of all transactions, mostly apartments. The NVM database registers ten house types which are assembled into four categories, namely: detached, semi-detached, attached, and apartment. The distribution of transactions to house type can be seen in Table 6.1. Clearly the fraction of apartments is dominant and the influence of detached houses on the total price development is small.

The selling prices are a priori divided in different segments, depending on neighborhoods and house type. The Amsterdam data is ordered according to the existing division in neighborhoods. These neighborhoods can be recognized by their postal codes so that the NVM database can be ordered accordingly. From these about 350 neighborhoods we construct 10 different sub-regional districts which generates relatively homogeneous groupings of neighborhoods with respect to house price development. As a classification of house types we use the one in Table 6.1. This results in a 40-segments classification produced from four property types and ten sub-regional districts. We refer to chapter 5 for a more extensive treatment of the segmentation.

On the basis of these transactions a model is constructed, as explained in the previous section, in which for each transaction a price is estimated which is compared to the actual purchase price. When the actual price differs more than 80% (about 4 times the standard deviation of the model) from the value calculated with the model, transactions are excluded (229

transactions (0,7%)), because they are considered unreliable. This results in a final database with 31219 transactions over the period 1985 – 1999Q2, a loss of 30,3% (13561 transactions) compared to the original database.

6.5.3 Breda region

The Breda database contains 25,644 transactions covering the period January 1985 until October 1999. The number of NVM transactions in the Breda region is relatively high, about 65% of the total number of transactions. The Breda region contains selling prices of different municipalities: Baarle-Nassau, Breda, Chaam, Dongen, Dussen, Geertruidenberg, Gilze en Rijen, 's Gravenmoer, Made en Drimmelen, Nieuw-Ginniken, Oosterhout, Prinsenbeek, Raamsdonk, Teteringen en Waspik.

To be able to correct for differences in quality of the houses it is necessary to take into account housing characteristics like house size, lot size, year of construction, postal code, etc. This leads to extra demands on the data resulting in 21,175 usable transactions. Especially in the earlier years less than half of the database could be used.

The sales volume per year doubled during the period 1985 – 1999Q2. The fraction of sales in the municipality of Breda alone is more than 45% of all transactions. The NVM database registers ten house types which are assembled into four categories, namely: detached, semi-detached, attached, and apartment. The distribution of transactions to house type can be seen in Table 6.1. Clearly the fraction of single-family homes is dominant.

The selling prices are a priori divided in different segments, depending on neighborhoods and house type. We distinguishes 4 different sub-regional districts which generate relatively homogeneous groupings of neighborhoods with respect to house price development. As a classification of house types we use the one in Table 6.1. This results in a 16-segment classification produced from four property types and four sub-regional districts.

On the basis of these transactions a model is constructed, as explained in the previous section, in which for each transaction a price is estimated which is compared with the actual purchase price. When the actual price differs more than 60% (about 4 times the standard deviation of the model) from the value calculated with the model, transactions are excluded (50 transactions), because they are considered unreliable. This results in a final database with 21125 transactions over the period 1985 – 1999 October, a loss of 17,6% (4519 transactions) compared to the original database.

6.6 Model results: valuations

The model of selling prices in the Breda region is specified as described in equation (6.7). One general trend (μ_t) is specified as a random walk with drift, 4 district trends (ϑ_t), and finally 4 house type trends (λ_t), both as random walks. This results in 16 different trends. The districts are divided in 73 neighborhoods (ϕ) (postal area), for which we assume separate levels, modeled as random effects.

The model contains 50 coefficients of explanatory variables, and 5 variances to be estimated. The definitions of the variables are provided in Table 6.12 and 6.13 in the appendix. The additive variables, as explained in section 6.3, are specified as

$$\beta_1 \ln(\text{HouseSize800}+\beta_2\text{HouseSizeRest}+\beta_3\text{PlotSize500}+\beta_4\text{PlotSizeRest} \\ +\beta_5\text{GarageDetached}+\beta_6\text{GarageAttached}+\beta_7\text{GarageBuiltIn}).$$

The estimation results are shown in the Tables 6.2 and 6.3. In the appendix more results are shown: Table 6.16 contains the neighborhood levels with N the number of observations per neighborhood, and between brackets the standard deviation. Table 6.15 contains the coefficients for the different housing types, and Table 6.14 provides the coefficients for the interior and exterior maintenance (between brackets standard deviations are provided).

All coefficients have the correct sign. An increase of the House size by 10% leads to an increase of the value by approximately $0.673 \times 10\% \simeq 7\%$. The coefficient for a detached garage is somewhat lower than the other garage coefficients. Maybe this is due to the fact that the detached garages are more common in the rural areas than in the city. A listed building is about 15% more expensive than a "normal" house. The linear drift has a coefficient of 0.0066, indicating an average yearly price rise of $12 \times 0.0066 \simeq 8\%$ over the whole period.

Other interesting findings with respect to age categories are, that the value of houses built before 1900 is 5.5% higher than that of houses built between $1900 - 1920$. The value of houses built between $1920 - 1945$ is 4.7% higher than that of houses built between $1900 - 1920$. For houses built after 1945 the value diminishes 0,5% with age (selling year minus year of construction). An increase of the selling period in days causes a decrease of the selling price 1.4% per week.

The coefficients for interior and exterior maintenance in Table 6.14 show differences of respectively 0.26 (30%) and 0.23 (26%) between perfect and poor maintenance.

The standard deviations for the trends and the measurement equation (6.7) are provided in Table 6.3. The standard deviation of the measurement equation σ is 0.1262, which can be interpreted as a standard deviation of a valuation for an individual dwelling. So, 66 percent of the residuals are within one standard deviation. The standard deviations for the random walks, the general trend (μ), the district trends (ϑ), and the house type trends (λ), are small compared to the standard deviation of the measurement equation. The general price deviation per year,

Table 6.2: Estimation results Breda region (HTM).

Variable	Coefficient
HouseSize800	0.673 (0.0057)
HouseSizeRest	0.883 (0.0423)
PlotSize500	0.901 (0.0219)
PlotSizeRest	0.085 (0.0033)
GarageDetached	45.82 (2.3922)
GarageAttached	70.07 (3.1811)
GarageBuiltIn	55.35 (4.3825)
NRooms	0.0134 (0.00104)
Age1900	-0.1745 (0.0089)
Age1920	-0.2280 (0.0064)
Age1945	-0.1817 (0.0046)
Age	-0.0052 (0.00012)
Listed	0.1385 (0.0197)
Term	-0.0020 (0.0003)
SalesConditions	-0.0019 (0.0132)
LivingRoom1	0.0342 (0.0027)
LivingRoom2	0.0214 (0.0073)
LivingRoom3	0.0251 (0.0052)
LivingRoom4	0.0067 (0.0026)
LivingRoom5	0.0075 (0.0051)
Time in months (κ)	0.0066 (0.0006)

Table 6.3: Estimation results standard deviations (HTM).

Region	Breda	Amsterdam
σ	0.1262	0.1824
$\sigma\sqrt{q_1}$ (μ)	0.0074	0.0116
$\sigma\sqrt{q_2}$ (θ)	0.0060	0.0110
$\sigma\sqrt{q_3}$ (λ)	0.0024	0.0630
$\sigma\sqrt{q_4}$ (ϕ)	0.0983	0.1312

apart from the drift, has a standard deviation of $\sqrt{12} \times 0.0074 \simeq 2.6\%$. The standard deviation of the random effects for the neighborhoods is about 10%. This means that a neighborhood level is in 66 percent of the cases within -10% and $+10\%$ from the district level.

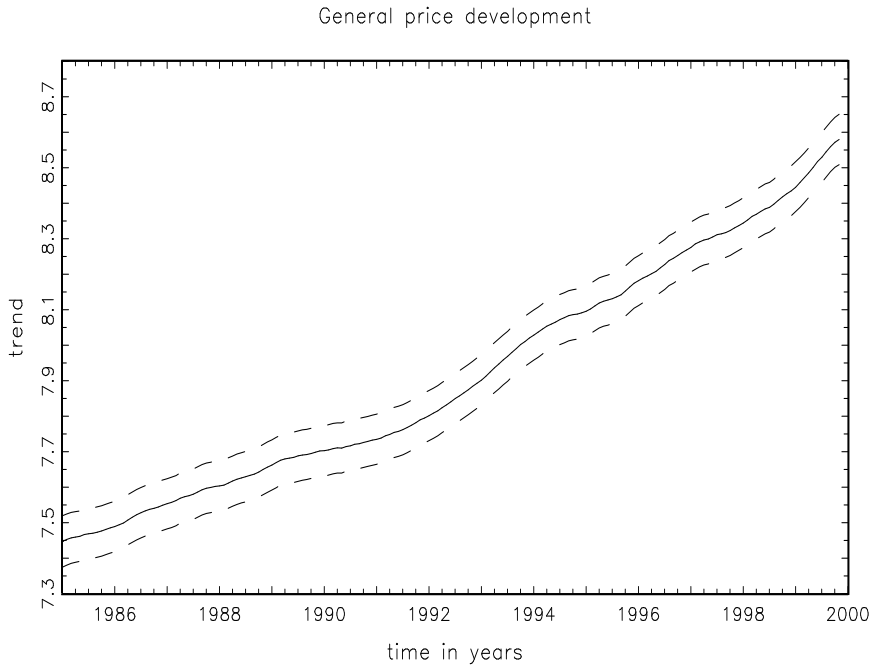

Figure 6.1: General trend for the Breda region on a monthly basis (HTM).

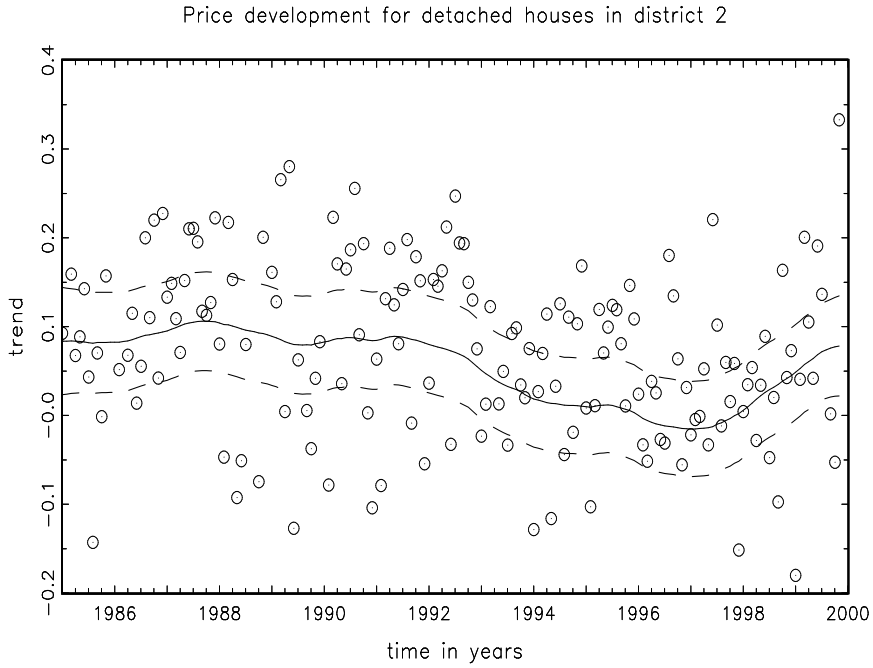

Figure 6.2: A specific cluster trend for the Breda region on a monthly basis (HTM).

Figure 6.1 gives the general trend for the region as a whole while Figure 6.2 displays an example of the trend for a specific cluster (semi-detached in a certain district; number of observations per month varies from a minimum of 3 and a maximum of 58) as a deviation from the general trend. The y axis is in logarithms, so an increase of .1 means that the increase in selling prices is approximately 10%. The dashed lines indicate the 95% confidence intervals and the points the average standardized selling prices, corrected for individual characteristics and the general trend, see equation (6.7).

The model specification for the Amsterdam region almost coincides with the specification for the Breda region. In Table 6.3 the estimates of the variance for the Amsterdam region are shown. The standard deviation for the measurement equation σ is 18%, about 6% points more than in the Breda region.

6.7 Model results: price indices

In this section we compare the price indices obtained by the HTM to price indices based on a standard hedonic method as well as based on a simple weighted method of median selling prices. In the first subsection we describe these methods, in the next subsection we compare the results of all methods for the Amsterdam and Breda region. The last subsection deals with the reliability of the price indices.

Of course there is a number of more elaborate models the HTM can be compared to, for example the model provided by Schwann (1998). This model can be seen as a special case of the HTM; a model with a general trend specified as a random walk with drift, without local and house type trends. This comes down to zero variance for the local and house type trends in the HTM, i.e. $q_2 = q_3 = 0$. For the Breda model a likelihood ratio test is performed. The loglikelihood in this model equals 15748.2. For the alternative model with $q_2 = q_3 = 0$ the loglikelihood is 15549.5, so the difference in loglikelihood is almost 199. This means that the alternative model is rejected.

6.7.1 Comparison with simple-weighted and standard hedonic methods

We compare the price index form the HTM to a simple weighted price index because the latter is the one reported by the national brokerage organization. In the simple-weighted method the median selling price is calculated in period t and $t+1$ for every market segment. Next, a weighted average of the segment medians is calculated with the relative number of sales in the segment as weights, for both periods. The relative difference between the two weighted averages provides the price index. The weights are not fixed but are presented by the relative number of sales in each separate period (rolling basis).

In formula, with $i = 1, \ldots, B$, the market segments and t the period,

$M_{i,t}$ the median selling price in market segment i and period t,
M_t the weighted median selling price in period t,
$n_{i,t}$ the number of sales in market segment i and period t,
n_t the number of sales in period t.
Then

$$n_t = n_{1,t} + \cdots + n_{B,t},$$
$$M_t = (n_{1,t} \times M_{1,t} + \cdots + n_{B,t} \times M_{B,t})/n_t.$$

So the relative price movement equals $(M_{t+1}/M_t - 1) \times 100\%$.

In the HTM hedonic index the market segment price movements are constructed from the model, as described in section 6.4. Further, fixed weights are used to obtain a representative general price index from the segment price movements. The fixed weights in time are taken as the relative number of selling prices per market segment over a long reference period (1985 − 1999Q2). The model simultaneously provides trends for different districts and house types on a monthly basis. From these trends it is quite easy to construct price indices on a monthly, quarterly, or yearly basis. In the model corrections are also made for differences in structural and locational characteristics in order to be able to compare the selling prices.

The differences between both methods are quite obvious. The HTM hedonic index corrects for differences in characteristics of the houses, and the simple-weighted index does not correct for any difference. The HTM hedonic index uses fixed weights per period, the simple-weighted index has time varying weights.

The second comparison of price rates is between the HTM and a standard hedonic time dummy model. This approach incorporates additional time dummy variables into a regression covering more than one time period, the coefficients of such dummies reflecting the change in price from one period to another, see for example Case and Quigley (1991). The standard hedonic model (SHM) calculates a price index for a specific market segment (semi-detached houses in a specific district). The dependent variable and the functional form of the explanatory variables (52, including 24 dummy variables of neighborhoods, and 7 dummy variables for house types) are a subset of the variables in the HTM, due to this specific market segment. In the standard hedonic model the trend specifications are replaced by 60 time dummy variables.

The estimation results for the standard hedonic model with 4362 observations are provided in Table 6.4. Table 6.17 in the appendix contains the neighborhood levels within this segment. The standard deviation of regression and the coefficient of determination are reasonably well.

Note that not all coefficients are significant due to the smaller number of observations. In the next subsection the price index based on the coefficients of the time dummy variables is compared to the price index of the HTM.

Table 6.4: Estimation results for specific market segment in the Breda region (SHM).

Variable	Coefficient	Variable	Coefficient
HouseSize800	0.700 (0.0114)	Term	-0.002 (0.0005)
HouseSizeRest	0.481 (0.1643)	SalesConditions	0.047 (0.0259)
PlotSize500	0.674 (0.0307)	HouseType10	-0.035 (0.0074)
PlotSizeRest	0.082 (0.0262)	HouseType12	0.070 (0.0051)
GarageDetached	56.388 (3.8994)	HouseType13	0.086 (0.0292)
GarageAttached	78.459 (4.3428)	HouseType16	0.114 (0.0184)
GarageBuiltIn	52.012 (6.1897)	HouseType17	0.187 (0.0149)
NRooms	0.011 (0.0022)	HouseType18	0.086 (0.0178)
Age1900	-0.181 (0.0235)	HouseType19	0.058 (0.0447)
Age1920	-0.234 (0.0160)	MI1	0.082 (0.0140)
Age1945	-0.182 (0.0094)	MI2	0.048 (0.0104)
Age	-0.005 (0.0003)	MI4	-0.063 (0.0280)
Listed	0.227 (0.0468)	MI5	-0.187 (0.0548)
LivingRoom1	0.037 (0.0259)	ME1	0.045 (0.0146)
LivingRoom2	0.003 (0.0045)	ME2	0.038 (0.0110)
LivingRoom3	0.010 (0.0130)	ME4	-0.084 (0.0288)
LivingRoom4	0.018 (0.0078)	ME5	-0.184 (0.0567)
LivingRoom5	0.002 (0.0049)		
σ	0.1048	R^2	0.9271

6.7.2 Price indices for the Amsterdam and Breda region

In this subsection price indices are shown for the Amsterdam and Breda region for the HTM, the simple-weighted and the standard hedonic method.

In general we find that the HTM hedonic price index varies over house type and district for the Amsterdam region, while for the Breda region the price development varies merely over house type.

Firstly, Figure 6.3 compares the yearly price changes of the HTM and the simple-weighted method for the Amsterdam region over the period $1985 - 1999Q2$. It is notable that differences between these methods are substantial, for instance up to 16 percent points in 1989, despite the fact that there were more than 1000 transactions per year. So, it seems to be very important to correct for differences in characteristics of the houses sold.

Similar results for the Breda region are shown in Table 6.5. For small market segments the differences are even more apparent. Table 6.6 displays the yearly price changes of apartments in a specific district in the Breda region, with N begin the number of observations in a year. The simple-weighted index is much more volatile than the HTM index, and provides unrealistic price changes, for example a price fall of 5.4% in 1987, and a price rice of 27.2% in 1988. So the simple method seems incapable to produce a reliable index for small market segments.

The second comparison is between price indices of the HTM and the standard hedonic model (SHM). Table 6.7 presents the quarterly price index results for a specific market segment

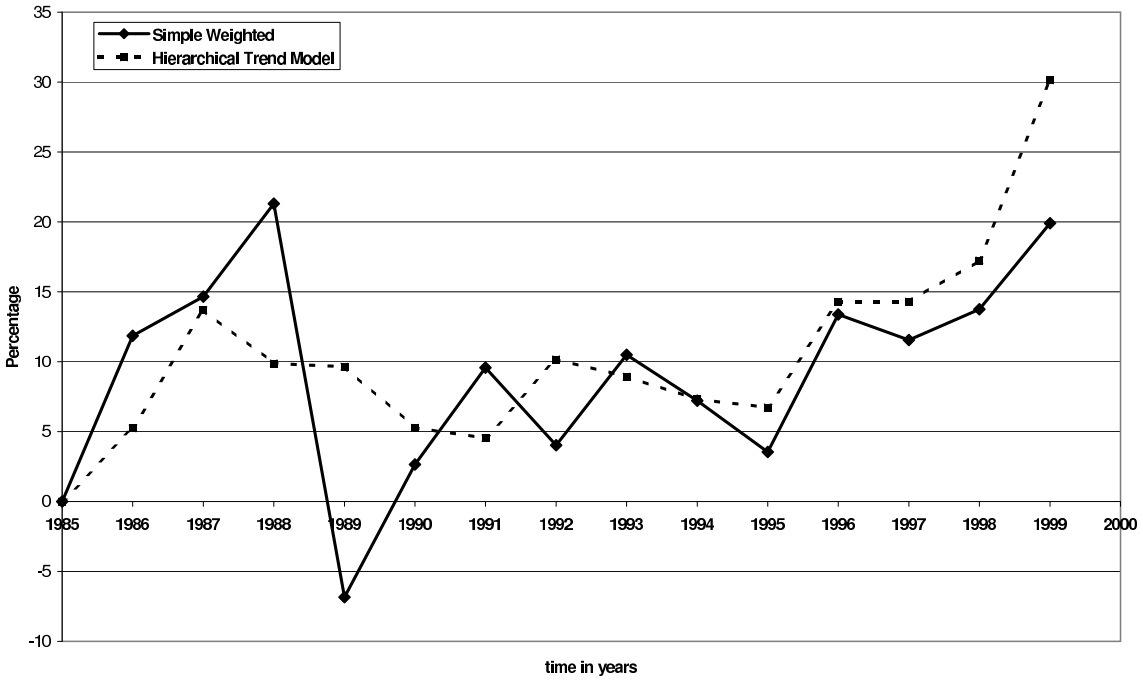

Figure 6.3: Price change for Amsterdam region.

Table 6.5: Price change in percentage per year for Breda region.

	Simple-weighted		HTM hedonic	
	Change	Cumulative	Change	Cumulative
1985				
1986	3.9	3.9	5.4	5.4
1987	7.3	11.4	6.1	11.9
1988	3.3	15.1	5.1	17.6
1989	7.2	23.4	6.3	25.0
1990	0.6	24.2	3.2	29.0
1991	4.9	30.3	4.9	35.3
1992	9.3	42.4	9.4	48.1
1993	10.8	57.8	12.7	67.0
1994	7.9	70.2	10.8	85.0
1995	9.0	85.6	7.2	98.3
1996	9.4	102.9	9.7	117.6
1997	7.3	117.8	8.5	136.1
1998	7.8	134.7	8.5	156.3
1999 Oct.	13.4	166.0	13.4	190.7

Table 6.6: Price Changes per year in percentages for small market segment in Breda region.

Year	N	Simple-weighted	HTM hedonic
1985	21	-	
1986	23	0.9	7.1
1987	18	-5.4	6.0
1988	22	27.2	4.8
1989	29	-3.1	6.5
1990	37	3.0	1.9
1991	39	7.9	3.9
1992	47	8.2	8.8
1993	53	17.8	12.5
1994	67	4.9	10.8
1995	91	9.2	6.7
1996	100	13.3	9.3
1997	115	1.9	8.8
1998	129	9.8	8.9
1999 Oct.	106	15.7	13.5

(semi-detached houses in a specific district), with N being the number of observations and the standard deviations between brackets. Both methods provide a different picture at this market level. The SHM price changes are more volatile, especially when few observations are available (for instance 1990.Q1, and 1990.Q2). This is reflected in the larger standard deviations for this method, about two times larger than the HTM standard deviations. So, the HTM index for small market segments on a quarterly basis seems more reliable than the index of the standard hedonic model.

6.7.3 Reliability

In the last subsection it was shown that the simple-weighted, and the standard hedonic method seemed to produce less reliable indices compared to the HTM. The reliability depends merely on the number of observations and the heterogeneity of the houses sold. The number of observations is dependent of the number of clusters, and the time period. The more clusters are distinguished (the smaller the market segment), and the shorter the time period considered, the less observations are available. In this subsection standard deviations are provided for the price changes in the Breda region. The price changes are produced on a monthly (M), quarterly (Q), and yearly (Y) basis, for the region as a whole, and for an "average" market segment (district and house type). Table 6.8 shows the standard deviations for the three methods in percentages.

The differences between the three methods are striking. The standard deviation for the HTM hedonic method is 2 to 6 times smaller than for the other methods. For instance, if a monthly submarket price change of 10% is computed by the simple-weighted index, the 95%

Table 6.7: Quarterly price changes for detached houses in district 3 of Breda region.

Q	N	Δ(%)HTM	Δ(%)SHM	Q	N	Δ(%)HTM	Δ(%)SHM
1985-2	34	0.95 (1.02)	1.57 (2.84)	1992-3	53	2.97 (0.79)	4.44 (1.90)
1985-3	40	0.85 (0.90)	0.43 (2.46)	1992-4	57	2.95 (0.81)	4.16 (2.01)
1985-4	35	1.05 (0.88)	0.10 (2.45)	1993-1	73	2.95 (0.79)	3.01 (1.86)
1986-1	40	1.01 (0.87)	0.15 (2.45)	1993-2	79	3.39 (0.79)	2.77 (1.71)
1986-2	49	3.03 (0.87)	3.39 (2.26)	1993-3	88	3.42 (0.76)	1.77 (1.63)
1986-3	34	1.90 (0.85)	3.38 (2.35)	1993-4	81	3.28 (0.76)	4.77 (1.63)
1986-4	26	0.87 (0.87)	-2.15 (2.74)	1994-1	87	2.68 (0.75)	3.09 (1.63)
1987-1	42	1.20 (0.87)	1.02 (2.63)	1994-2	81	2.28 (0.75)	2.15 (1.79)
1987-2	47	1.32 (0.87)	5.28 (2.24)	1994-3	61	1.68 (0.76)	2.09 (1.79)
1987-3	43	1.53 (0.87)	0.88 (2.22)	1994-4	69	0.33 (0.78)	-1.29 (1.85)
1987-4	31	0.97 (0.90)	1.03 (2.49)	1995-1	101	1.73 (0.75)	1.48 (1.65)
1988-1	39	0.59 (0.87)	-0.80 (2.54)	1995-2	97	2.10 (0.73)	3.13 (1.50)
1988-2	38	1.52 (0.88)	3.60 (2.41)	1995-3	82	1.17 (0.71)	1.42 (1.58)
1988-3	41	0.72 (0.87)	-2.60 (2.39)	1995-4	92	3.20 (0.73)	1.96 (1.60)
1988-4	43	2.57 (0.87)	3.32 (2.32)	1996-1	108	1.88 (0.70)	2.48 (1.49)
1989-1	55	3.14 (0.85)	4.60 (2.15)	1996-2	109	2.31 (0.70)	3.66 (1.43)
1989-2	37	1.46 (0.85)	3.24 (2.24)	1996-3	109	3.05 (0.69)	1.71 (1.43)
1989-3	42	0.41 (0.84)	-2.32 (2.38)	1996-4	103	2.30 (0.70)	4.31 (1.45)
1989-4	56	0.69 (0.86)	-0.84 (2.15)	1997-1	113	3.26 (0.68)	1.85 (1.44)
1990-1	35	1.21 (0.85)	6.57 (2.28)	1997-2	123	1.72 (0.67)	2.59 (1.37)
1990-2	37	0.18 (0.84)	-4.11 (2.49)	1997-3	111	1.17 (0.67)	0.17 (1.38)
1990-3	50	0.21 (0.83)	0.31 (2.29)	1997-4	135	1.13 (0.67)	0.46 (1.35)
1990-4	52	0.56 (0.80)	3.41 (2.09)	1998-1	153	2.27 (0.66)	3.33 (1.25)
1991-1	69	1.08 (0.80)	-3.37 (1.94)	1998-2	132	2.38 (0.66)	1.86 (1.25)
1991-2	70	1.62 (0.80)	4.44 (1.80)	1998-3	142	2.18 (0.65)	2.42 (1.27)
1991-3	60	1.30 (0.80)	-1.36 (1.86)	1998-4	134	2.58 (0.66)	4.02 (1.27)
1991-4	69	2.02 (0.81)	6.53 (1.87)	1999-1	136	3.24 (0.65)	1.10 (1.28)
1992-1	73	1.75 (0.79)	-0.07 (1.78)	1999-2	123	4.89 (0.66)	5.31 (1.31)
1992-2	74	2.63 (0.79)	0.93 (1.75)	1999-3	119	4.89 (0.69)	6.60 (1.35)
				Cum.		201.33	202.41

Table 6.8: Standard deviation of price changes for three methods.

	Simple-weighted			SHM			HTM		
Time period	Y	Q	M	Y	Q	M	Y	Q	M
Region	0.6%	1.2%	2.1%	0.44%	1.0%	1.4%	0.36%	0.6%	0.9%
Market segment	2.5%	5.0%	8.7%	0.85%	1.3%	2.4%	0.85%	1.2%	1.4%

confidence interval is provided by [-7.4%;27.4%]; the standard hedonic model produces a 95% confidence interval of [5.2;14.8%]. If it is computed by the HTM hedonic index this confidence interval is [7.2%;12.8%]. For a yearly regional price change these intervals are [8.8%;11.2%], [[9.1%;10.9%] and [9.3%;10.7%], respectively.

It can be concluded that the HTM hedonic index is far more accurate. Thus it is possible to obtain reliable indices on a more detailed level, and for small time periods, and therefore it is also more up-to-date. The differences between the HTM and the standard hedonic are more pronounced for monthly time periods and smaller market segments.

6.8 Model extensions

6.8.1 Time

The HTM as described in section 6.4, could easily be extended in several ways. We discuss some temporal and spatial modifications.

In the specification of the HTM for the general trend we propose a random walk with drift. The random walk with drift can be further generalized by allowing κ to vary over time:

$$\mu_{t+1} = \kappa_t + \mu_t + \eta_t, \qquad \eta_t \sim N(0, \sigma_\eta^2), \tag{6.13}$$

$$\kappa_{t+1} = \kappa_t + \zeta_t, \qquad \zeta_t \sim N(0, \sigma_\zeta^2), \tag{6.14}$$

with known μ_1, κ_1, and independent η and ζ. Equation (6.4) with these specifications for the trend is called the *local trend* model. The trend μ becomes smoother when we decrease σ_η^2; in the limiting case of $\sigma_\eta^2 = 0$, μ_t is said to follow an *integrated random walk*, since its first difference follows a random walk.

In fact, all specifications that can be put in state-space format, like all ARIMA models, can be estimated by the Kalman filter. So various trend specifications for the general and cluster trends can be used.

Another obvious generalization would be to vary the regression parameters β over time, for example $\beta_{t+1} = \beta_t + \varrho_t$, with $\varrho_t \sim N(0, q_5\sigma^2 I)$.

6.8.2 Space

From spatial econometrics two notions are known, spatial heterogeneity and spatial dependence, see for example Anselin (1988). Spatial heterogeneity can be described as follows: functional forms and parameters vary with location and are not homogeneous throughout the data set. And spatial dependence: the variation is a function of distance.

Spatial models for housing prices can be specified on an individual level (observation) and on a cluster level, for example neighborhood, or city level. Spatial models on an individual level

are complex to evaluate. Examples of such spatial models are provided by for example Can (1992), Dubin (1992) and Dubin (1998), and more recently by Wolverton and Senteza (2000). An example of a spatio-temporal model is given in Pace et al. (1998). Those models are not considered here. A drawback of the cluster level approach is that there might be undesirable discontinuities on borders, and it requires knowledge of the spatial structure, which might be different from available administrative clusters.

In the HTM the constants vary over time and neighborhood, so the HTM has spatial heterogeneity. The neighborhood levels are specified as random effects. Another possibility would be to specify the neighborhood levels as fixed effects. A drawback of the fixed effects specification is that nothing can be said about neighborhoods not included in the sample, for example neighborhoods without selling prices, and that it is more sensitive for outliers.

The spatial heterogeneity can be extended to other regression variables X_t. This means that β would vary over the different clusters. An example of such a specification is provided by

$$\beta_j = \beta + \varsigma_j, \qquad \varsigma_j \sim N(0, \sigma_\varsigma^2 I),$$

with j indicating cluster j.

Another spatial extension would be to model for the districts the initial levels ϑ_0 as

$$\vartheta_0 | \pi \sim N(V\pi, \sigma^2 \Psi), \tag{6.15}$$

where V contains explanatory variables for district value levels. Examples are crime rate, and distance from the city center. This is an example of a hierarchical model. This kind of models is described by Can (1992), Francke (1996) and Orford (1999). For a thorough treatment of hierarchical, or multilevel models we refer to Bryk and Raudenbush (1992), Goldstein (1995), Longford (1993), and O'Hagan (1994).

Spatial dependence could be introduced by the variance matrix, which is specified as a spatial autocorrelation matrix. An example of a spatial autocorrelation matrix is provided by $\Psi = (I_B - \rho W)'(I_B - \rho W)$. with matrix elements w_{ij} defined by

$$w_{ij} = \begin{cases} 1 & \text{cluster } i \text{ and } j \text{ are adjacent,} \\ 0 & \text{otherwise.} \end{cases}$$

With a scaling factor, we can use the same matrix to model correlations in the district trend disturbances; in equation (6.9), $\omega_t \sim N(0, q_3 \sigma^2 \Psi)$. These more elaborate specifications are especially valuable if in some districts few observations are available. With this model we can also value houses in districts without any selling price data.

6.8.3 Modification of the HTM

In this subsection some modifications of the HTM as specified in section 6.4 are presented. Firstly, in equation (6.11) it is assumed that β is constant for all property types. In this subsection we allow for a market segmentation of apartments and single-family houses, so β varies over both these categories. Secondly, in the same equation homoskedasticity was assumed over the four property types. In the modified HTM we assume heteroskedasticity. Finally, in equations (6.8) and (6.9) it was assumed that the variances are constant over the different market segments. Now we assume the variances to vary over districts and property types.

Table 6.9: Estimation results Breda region (Modified HTM).

Variable	Single-family	Apartments
HouseSize800	0.675 (0.0060)	0.489 (0.0188)
HouseSizeRest	0.842 (0.0478)	
PlotSize500	0.814 (0.0206)	
PlotSize1000	0.240 (0.0335)	
PlotSizeRest	0.078 (0.0041)	
GarageDetached	43.090 (2.2792)	
GarageAttached	69.776 (3.0884)	
GarageBuiltIn	40.776 (4.6074)	
Garage		93.282 (10.640)
NRooms	0.014 (0.0010)	0.025 (0.0058)
Age1900	-0.153 (0.0091)	-0.264 (0.0714)
Age1920	-0.206 (0.0064)	-0.645 (0.0524)
Age1945	-0.163 (0.0046)	-0.557 (0.0402)
Age	-0.005 (0.0001)	-0.008 (0.0004)
Listed	0.132 (0.0200)	
Term	-0.002 (0.0003)	-0.003 (0.0015)
SalesConditions	-0.006 (0.0127)	
LivingRoom1	0.032 (0.0027)	0.060 (0.0109)
LivingRoom2	0.016 (0.0071)	0.058 (0.0354)
LivingRoom3	0.020 (0.0049)	0.104 (0.0390)
LivingRoom4	0.008 (0.0026)	0.018 (0.0097)
LivingRoom5	0.011 (0.0052)	0.049(0.0166)

Tables 6.9 and 6.10 provide the estimation results for the modified HTM with between brackets the standard deviations. Note that the estimation results for β are different for apartments and single-family houses. Standard deviations are provided in Table 6.11. For the different property types the standard deviation of the measurement equation varies between 0.1073, and 0.1649, indicating heteroskedasticity. The overall standard deviation, defined as a weighted average of the standard deviations of the measurement equation per house type (with as weights the number of observations) is 0.1107, about one percent point less than the result of the HTM in Table 6.3. The standard deviations for the trends also vary over districts an house types.

Table 6.10: Estimation results Maintenance Breda region (Modified HTM).

Variable	Single-family	Apartments
MI1	0.091 (0.0064)	0.092 (0.0211)
MI2	0.058 (0.0042)	0.031 (0.0145)
MI4	-0.041 (0.0088)	-0.040 (0.0389)
MI5	-0.192 (0.0211)	-0.417 (0.1186)
ME1	0.060 (0.0066)	0.087 (0.0256)
ME2	0.047 (0.0044)	0.045 (0.00179)
ME4	-0.084 (0.0093)	-0.144 (0.0640)
ME5	-0.169 (0.0231)	

Table 6.11: Estimation results standard deviations (Modified HTM).

House type or district	1	2	3	4
σ	0.1073	0.1155	0.1649	0.1613
$\sigma\sqrt{q_1}$ (μ)	0.00765			
$\sigma\sqrt{q_2}$ (θ)	0.00353	0.00542	0.00074	0.00046
$\sigma\sqrt{q_3}$ (λ)	0.00366	0.00128	0.00781	0.00804
$\sigma\sqrt{q_4}$ (ϕ)	0.0801			

6.9 Conclusions

This chapter presented a dynamic hedonic price model for selling prices of houses. The model considered is a hierarchical trend model with general and cluster price trends. The clusters are constructed by location and house type. This model can be seen as an extension of a dummy variable model, with time varying constants for the different clusters. For the general trend a random walk with drift is assumed, for the cluster trends random walks are assumed. The coefficients of the explanatory variables are kept constant over time, location, and house type. These kind of dynamic models, even with varying coefficients, can be put in state-space format, so they can be estimated by the (diffuse) Kalman filter.

Model results are shown for the regional housing markets of Breda and Amsterdam as well as for local housing markets within these regions. It is shown that an estimate of the value of an individual house can be produced with an average standard deviation of 18% for the Amsterdam region, and 13% for the Breda region.

HTM hedonic price indices were constructed from the trends of the hierarchical trend model. These indices were compared with standard hedonic indices (SHM) and a simple weighted index published by the national brokerage organization. The question was which method measures the most adequate price change estimates for standardized houses of constant quality, thereby measuring price changes in the market due to market forces only. The findings of this research are summarized below.

In general using the HTM provides more up-to-date, detailed, and reliable results than using the SHM an the simple weighted method.

Table 6.12: Variable definitions Breda region.

Variable	Definition
HouseSize800	the minimum of the house size in cubic meters, and 800
HouseSizeRest	the maximum of the house size in cubic meters - 800, and 0
PlotSize500	the minimum of the lot size in square meters, and 500
PlotSizeRest	the maximum of the lot size in square meters - 500, and 0
GarageDetached	1 if detached garage, 0 otherwise
GarageAttached	1 if attached garage, 0 otherwise
GarageBuiltIn	1 if built-in garage, 0 otherwise
NRooms	number of rooms
Age1900	1 if year of construction $<$ 1900, 0 otherwise
Age1920	1 if 1900 $\leq$ year of construction $<$ 1920, 0 otherwise
Age1945	1 if 1920 $\leq$ year of construction $<$ 1945, 0 otherwise
Age	$\begin{cases}$ if year of construction $\geq$ 1945, sellingyear - year of construction, 0 otherwise
Listed	1 if listed building, 0 otherwise
Term	Sellingperiod in days
SalesConditions	1 of no legal charges, 0 otherwise
Time in months	selling date in months from 1 January 1985
MI	interior maintenance -1 Unknown,1 Excellent,2 Good,3 Reasonable,4 Moderate,5 Poor
ME	exterior maintenance -1 Unknown,1 Excellent,2 Good,3 Reasonable,4 Moderate,5 Poor
LivingRoom	Type of living Room -1 Unknown, 1 L-shaped Room, 2 T-shaped Room, 3 Z-shaped Room, 4 Through Room, 5 Room en suite

When small market segments with few transactions are concerned the use of the HTM appears to be the only accurate price index construction method, especially when indices are necessary on a monthly or quarterly basis with even less transactions per period. In that case both the SHM and simple weighted method produce less reliable results because of this small number of observations.

When studying yearly price development on both the regional market and the inner regional markets the simple method, though in general use, appears to be unreliable even on a yearly basis.

Table 6.13: Definition House types.

House Type	Description
10	Simple house
11	Middle class house
12	Mansion
13	Villa
14	Country house
15	Country Estate
16	Bungalow
17	Patio-bungalow
18	Semi-bungalow
19	Split level house
20	Meander house
21	Apartment on ground floor
22	Apartment on upper level
23	House with upper and lower levels
24	Apartment in building with a common entrance hall
25	House situated along a canal
26	Maisonette
27	Sheltered housing
28	Apartment with elevator
29	Apartment without elevator
30	House with ancillary office accommodation / house with surgery
31	Drive-in house
32	Converted farmhouse

Table 6.14: Estimation results Maintenance Breda region (HTM).

Variable	Coefficient	Standard Deviation	T-value
MI1	0.0907	0.0062	14.54
MI2	0.0526	0.0041	12.70
MI4	0.0468	0.0091	-5.13
MI5	-0.1760	0.0219	-8.05
ME1	0.0593	0.0065	9.06
ME2	0.0467	0.0044	10.67
ME4	-0.0887	0.0096	-9.20
ME5	-0.1717	0.0233	-7.37

Table 6.15: Estimation results House type Breda region (HTM).

Variable	Coefficient	T-value
HouseType10	-0.058 (0.0037)	-15.81
HouseType12	0.107 (0.0032)	33.29
HouseType13	0.236 (0.0064)	37.09
HouseType14	0.2690 (0.0100)	26.89
HouseType15	0.1441 (0.0496)	2.90
HouseType16	0.1757 (0.0083)	21.28
HouseType17	0.1979 (0.0102)	19.46
HouseType18	0.1657 (0.0065)	25.52
HouseType19	0.0191 (0.0226)	0.84
HouseType20	0.0223 (0.1267)	0.176
HouseType30	0.0567 (0.0199)	2.84
HouseType31	-0.0799 (0.0119)	-6.72
HouseType32	0.1068 (0.0113)	9.47
HouseType21	-0.1391 (0.0238)	-5.84
HouseType22	-0.1355 (0.0219)	-6.18
HouseType23	-0.2732 (0.0294)	-9.28
HouseType24	-0.0155 (0.0235)	-0.66
HouseType26	-0.1147 (0.0152)	-7.56
HouseType27	-0.2113 (0.0289)	-7.32
HouseType28	-0.0107 (0.0089)	-1.19
HouseType29	-0.0878 (0.0090)	9.71

Table 6.16: Estimation results neighborhood levels Breda region (HTM).

Neigh	Coefficient	N	Neigh	Coefficient	N
4835	0.068 (0.025)	764	4941	0.034 (0.020)	509
4836	0.050 (0.036)	19	4942	0.013 (0.021)	331
4837	0.118 (0.026)	226	4944	-0.024 (0.023)	37
4847	-0.041 (0.025)	517	5101	0.009 (0.021)	312
4851	0.061 (0.025)	379	5102	0.007 (0.021)	265
4856	-0.114 (0.048)	7	5103	0.030 (0.021)	366
4858	0.055 (0.061)	3	5104	0.019 (0.021)	312
4859	0.071 (0.078)	1	5105	-0.031 (0.044)	8
4902	-0.115 (0.025)	805	5106	-0.050 (0.031)	25
48181	0.018 (0.025)	298	5107	-0.093 (0.078)	1
48182	0.109 (0.026)	266	5109	-0.002 (0.022)	146
48183	0.005 (0.026)	207	5121	0.002 (0.020)	819
48184	0.138 (0.034)	24	5122	0.002 (0.022)	198
48185	0.087 (0.029)	59	5124	-0.030 (0.027)	40
48191	0.116 (0.026)	144	5165	-0.073 (0.022)	151
48192	0.058 (0.026)	240	48141	0.003 (0.021)	402
49040	-0.579 (0.056)	4	4812	0.060 (0.023)	492
49041	-0.104 (0.025)	743	4815	0.042 (0.024)	255
4811	0.068 (0.034)	460	4816	0.040 (0.024)	179
4813	-0.068 (0.034)	395	4825	-0.005 (0.049)	6
4817	-0.051 (0.033)	1074	4826	-0.052 (0.023)	492
4834	0.063 (0.033)	1004	4827	-0.038 (0.025)	168
4838	-0.001 (0.036)	57	4849	0.039 (0.024)	169
4839	0.015 (0.037)	44	4855	0.068 (0.029)	43
4841	0.018 (0.033)	844	4861	0.045 (0.026)	98
4854	0.040 (0.035)	404	4903	-0.075 (0.033)	22
48142	-0.084 (0.038)	36	4905	0.068 (0.026)	88
4271	-0.095 (0.023)	113	4906	0.032 (0.030)	40
4273	-0.057 (0.022)	175	4908	0.071 (0.024)	287
4822	-0.021 (0.020)	577	4909	-0.006 (0.027)	67
4823	0.041 (0.021)	392	4911	0.044 (0.028)	50
4824	-0.021 (0.020)	684	5111	-0.023 (0.026)	93
4901	0.100 (0.020)	952	5113	-0.137 (0.035)	19
4907	0.064 (0.020)	1448	5114	-0.154 (0.061)	3
4921	0.073 (0.021)	338	5125	-0.028 (0.037)	16
4924	0.040 (0.030)	27	5126	0.004 (0.024)	239
4931	0.061 (0.021)	492			

Table 6.17: Estimation result neighborhood levels for specific market segment in Breda region (SHM).

Neigh	Coefficient	Neigh	Coefficient
4873	0.073 (0.024)	5101	0.141 (0.023)
4822	0.117 (0.021)	5102	0.129 (0.021)
4823	0.181 (0.021)	5103	0.159 (0.022)
4824	0.105 (0.021)	5104	0.133 (0.022)
4901	0.205 (0.021)	5105	0.207 (0.080)
4907	0.188 (0.020)	5106	0.147 (0.064)
4921	0.208 (0.020)	5109	0.130 (0.024)
4924	0.117 (0.055)	5121	0.111 (0.021)
4931	0.159 (0.021)	5122	0.124(0.023)
4941	0.162 (0.021)	5124	0.155(0.052)
4942	0.140 (0.027)	5165	0.077(0.024)
4944	0.112 (0.021)	48141	0.087(0.023)

Appendix A

A.1 Derivation of marginal likelihood

Consider the linear model as provided in section 2.2.1. Define the $(n \times m)$ matrix A such that $A'X = 0$ and $[A\ X]$ has rank n, where $m = (n-k)$. The density $f(A'y|\theta, \sigma^2)$ is provided by

$$\begin{aligned} f(A'y|\theta,\sigma^2) &= (2\pi\sigma^2)^{-m/2}|A'\Omega A|^{-1/2}\exp\left\{-\frac{1}{2\sigma^2}y'A(A'\Omega A)^{-1}A'y\right\} \\ &= (2\pi\sigma^2)^{-m/2}|A'A|^{-1/2}|X'X|^{1/2}\left|X'\Omega^{-1}X\right|^{-1/2}|\Omega|^{-1/2}\exp\left\{-\frac{1}{2\sigma^2}y'\Omega^{-1}M_X^{\Omega}y\right\}. \end{aligned}$$

Proof: define $B = \begin{bmatrix} \Omega A & X \end{bmatrix}$. Then $B'A(A'\Omega A)^{-1}A' = \begin{bmatrix} A & 0 \end{bmatrix}'$ and $B'\Omega^{-1}M_X^{\Omega} = \begin{bmatrix} A & 0 \end{bmatrix}'$, so $A(A'\Omega A)^{-1}A' = \Omega^{-1}M_X^{\Omega}$. This implies that $y'A(A'\Omega A)^{-1}A'y$ provides the residual sum of squares, and for $\Omega = I_n$ that $A(A'A)^{-1}A' = M_X$.

Further,

$$\begin{aligned} |\Omega|^{1/2}|A'A|^{1/2}|X'X|^{1/2} &= |\Omega|^{1/2}\begin{vmatrix} A & X \end{vmatrix} = \begin{vmatrix} A'\Omega A & A'\Omega X \\ X'\Omega A & X'\Omega X \end{vmatrix}^{1/2} \\ &= |A'\Omega A|^{1/2}\left|X'\Omega X - X'\Omega A(A'\Omega A)^{-1}A'\Omega X\right|^{1/2} \\ &= \left|X'\Omega X - X'M_X^{\Omega}\Omega X\right|^{1/2} = |X'X|\left|X'\Omega^{-1}X\right|^{-1/2}, \end{aligned}$$

so

$$|A'\Omega A| = |\Omega|\,|A'A|\,|X'X|^{-1}\left|X'\Omega^{-1}X\right|,$$

leading to the mentioned result.

A.2 Computation of marginal likelihood in the ARX(1) model

In this appendix explicit expressions for the marginal likelihood in the autoregressive model (2.1) - (2.3) are derived by GLS. This model can be rewritten as

$$\begin{aligned} y_1 &= \mu + x_1\beta + u_1, \\ y_t - \rho y_{t-1} &= (1-\rho)\mu + (x_t - \rho x_{t-1})\beta + \varepsilon_t,\ t = 2, \ldots, n. \end{aligned}$$

Define

$$\begin{array}{lllllll} y(\rho) & = & \left(\begin{array}{cccc} y_1 & y_2(\rho) & \cdots & y_n(\rho) \end{array} \right)', & & y_t(\rho) & = & y_t - \rho y_{t-1}, \\ x_t^*(\rho) & = & x_t^* - \rho x_t^*, & & X^*(\rho) & = & \left(\begin{array}{ccccc} x_1^* & x_2^*(\rho)' & \cdots & x_n^*(\rho)' \end{array} \right)', \\ \beta^* & = & \left(\begin{array}{cc} \mu & \beta' \end{array} \right)', & & x_t^* & = & \left(\begin{array}{cc} 1 & x_t \end{array} \right), \\ \mathrm{Var}(\omega) & = & \Psi = \mathrm{diag}(\begin{array}{ccc} (1-\rho^2)^{-1} & 1 \cdots 1 \end{array}), & & \omega & = & \left(\begin{array}{ccc} u_1 & \varepsilon_2 \cdots \varepsilon_n \end{array} \right). \end{array}$$

The GLS estimator of β^* is provided by $\widehat{\beta}^* = \left(X^*(\rho)'\Psi^{-1}X^*(\rho)\right)^{-1} X^*(\rho)'\Psi^{-1}y(\rho)$, with

$$\begin{aligned} X^*(\rho)'\Psi^{-1}X^*(\rho) &= (1-\rho)\begin{pmatrix} n-(n-2)\rho & \Sigma_{x,\rho} \\ \Sigma_{x,\rho}' & (1-\rho)^{-1}\Sigma_{xx,\rho} \end{pmatrix}, \\ X^*(\rho)'\Psi^{-1}y(\rho) &= \begin{pmatrix} (1-\rho)\Sigma_{y,\rho} & \Sigma_{xy,\rho}' \end{pmatrix}'. \end{aligned}$$

It follows that

$$\begin{aligned} \ln\left|X'\Omega^{-1}X\right| &= \ln\left|X^*(\rho)'\Psi^{-1}X^*(\rho)\right| \\ &= \ln(n-(n-2)\rho)(1-\rho) + \ln\left|\Sigma_{xx,\rho} - \frac{(1-\rho)}{n-(n-2)\rho}\Sigma_{x,\rho}'\Sigma_{x,\rho}\right|, \\ \widehat{\mu} &= F\left(\Sigma_{y,\rho} - \Sigma_{x,\rho}\Sigma_{xx,\rho}^{-1}\Sigma_{xy,\rho}\right), \text{ and} \\ \widehat{\beta} &= \Sigma_{xx,\rho}^{-1}\Sigma_{xy,\rho} - (1-\rho)\Sigma_{xx,\rho}^{-1}\Sigma_{x,\rho}'\widehat{\mu}, \end{aligned}$$

with $F^{-1} = \left(n-(n-2)\rho-(1-\rho)\Sigma_{x,\rho}\Sigma_{xx,\rho}^{-1}\Sigma_{x,\rho}'\right)$. Note that $\ln|\Omega| = -\ln(1-\rho^2)$, and $\ln|X'X| = \ln(n) + \ln|\tilde{X}'\tilde{X}|$. The residual sum of squares is provided by

$$\mathrm{RSS}_{\mu,\beta}(\rho) = \Sigma_{yy,\rho} - (1-\rho)\Sigma_{y,\rho}\widehat{\mu} - \Sigma_{xy,\rho}'\widehat{\beta}.$$

Collecting terms leads to the mentioned result (2.18).

A.3 Priors coherent with marginal likelihood

Lemma Consider the linear model as provided in section 2.2.1. Define $y^* = A'y/\sqrt{y'M_Xy}$, for definitions of A and M_X see also section 2.2.1.

If $f(\theta|y) = f(\theta|y^*)$, and $\pi(\beta, \sigma^2|\theta) \sim NIG(a(\theta), d(\theta), m(\theta), V(\theta))$, so

$$\pi(\beta, \sigma^2|\theta) \propto \sigma^{-(d(\theta)+k+2)} \exp\left\{-\frac{(\beta - m(\theta))\prime V(\theta)^{-1}(\beta - m(\theta)) + a(\theta)}{2\sigma^2}\right\},$$

then $\pi(\sigma^2, \beta) \propto \sigma^{-2}$.

Proof: $f(\theta|y) = f(\theta|y^*)$ implies that $f(y|\theta) \propto f(y^*|\theta)$. The marginal likelihood $f(y^*|\theta)$ is provided by $f(y^*|\theta) \propto |X'\Omega^{-1}X|^{-1/2} |\Omega|^{-1/2} \left(y'\Omega^{-1}M_X^{\Omega}y\right)^{-\frac{n-k}{2}}$.
$f(y|\theta)$ can be expressed as

$$\begin{aligned}
f(y|\theta) &= \int\int f(y|\beta, \sigma^2, \theta)\pi(\beta, \sigma^2|\theta)d\beta d\sigma^2 \\
&\propto \int\int \sigma^{-(n+d(\theta)+k+2)} |\Omega|^{-1/2} \\
&\times \exp\left\{-\frac{1}{2\sigma^2}(y - X\beta)'\Omega^{-1}(y - X\beta)\right\} \\
&\times \exp\left\{-\frac{1}{2\sigma^2}\left[(\beta - m(\theta))' V(\theta)^{-1}(\beta - m(\theta)) + a(\theta)\right]\right\} d\beta d\sigma^2 \\
&\propto |V^*|^{1/2} |\Omega|^{-1/2} \left(y'\Omega^{-1}y - m^{*\prime}(V^*)^{-1} m^* + a(\theta) + m(\theta)V(\theta)^{-1}m(\theta)\right)^{-\frac{n+d(\theta)}{2}},
\end{aligned}$$

where $V^* = (X'\Omega^{-1}X + V(\theta)^{-1})^{-1}$ and $m^* = V^*(X'\Omega^{-1}y + V(\theta)^{-1}m(\theta))$. It follows that $f(y|\theta) \propto f(y^*|\theta)$ only for $a(\theta) = 0$, $d(\theta) = -k$, $V(\theta)^{-1} \to 0$.■

A.4 Adaptation of the diffuse Kalman filter

In this appendix the recursions of the diffuse Kalman filter are provided for state-space models as described in section 4.2. For a detailed description of the diffuse Kalman filter, see De Jong (1991a) and De Jong (1991b). The formulae for the diffuse Kalman filter for $t = 1, \ldots, T$ are

$$\begin{array}{llll} F_t & = & Z_t P_t Z_t' + H_t, & K_t & = & T_t P_t Z_t' F_t^{-1} \\ L_t & = & T_t - K_t Z_t, & M_t & = & H_t - K_t Z_t, \\ V_t & = & (y_t, 0) - Z_t A_t, & S_{A,t} & = & S_{A,t-1} + V_t' F_t^{-1} V_t, \\ A_{t+1} & = & T_t A_t + K_t V_t, & P_{t+1} & = & T_t P_t L_t' + R_t Q_t R_t', \end{array}$$

with initial conditions $A_1 = (a_0, A_0)$, $P_1 = R_0 Q_0 R_0'$, and $S_{A,0} = 0$.

Define $S_{A,T} = \begin{pmatrix} q & -s' \\ -s & S \end{pmatrix}$, with q a scalar. The diffuse likelihood is provided by

$$-2\ell_D(\theta, \psi, \sigma^2) = m \ln\left(2\pi\sigma^2\right) + \sum_{t=1}^{T} |F_t| + \ln|S| + \sigma^{-2}\left(q - s' S^{-1} s\right).$$

In order to evaluate the marginal likelihood $\ell_{M_\beta}(\theta, \sigma^2)$ some recursions need to be added, for $t = 1, \ldots, T$,

$$A_{t+1}^* = T_t A_t^*, \quad V_t^* = (y_t, 0) - Z_t A_t^*, \quad S_{A,t}^* = S_{A,t}^{*-1} + V_t^{*\prime} V_t^*,$$

with initial conditions $A_1^* = (a_0, A_0)$, and $S_{A,0}^* = 0$. Define $S_{A,T}^* = \begin{pmatrix} q^* & -s^{*\prime} \\ -s^* & S^* \end{pmatrix}$, with q^* a scalar, then the marginal likelihood is provided by

$$-2\ell_{M_\beta}(\theta, \sigma^2) = m \ln(2\pi\sigma^2) + \sum_{t=1}^{T} \ln|F_t| + \ln|S| - \ln|S^*| + \sigma^{-2}\left(q - s' S^{-1} s\right).$$

The difference with the diffuse likelihood is the term $\ln|S^*|$.

The marginal likelihood $\ell_{M_{\beta,\sigma}}$ is provided by

$$-2\ell_{M_{\beta,\sigma}}(\theta) = m\left(\ln\left(\frac{q - s' S^{-1} s}{q^* - s^{*\prime} S^{*-1} s^*}\right) + \ln(\pi)\right) - 2\ln\frac{1}{2}\Gamma(\frac{m}{2}) + \sum_{t=1}^{T} \ln|F_t| + \ln|S| - \ln|S^*|.$$

A.5 Multivariate t-distribution

The multivariate t-distribution $f_n(x|\omega, \Omega, v)$ is provided by

$$f_n(X|\omega, \Omega, v) = \frac{v^{v/2}\Gamma((v+n)/2)}{\pi^{n/2}\Gamma(v/2)} |\Omega|^{-1/2} \left[v + (x-\omega)'\Omega^{-1}(x-\omega)\right]^{-(n+v)/2},$$

where $v > 0$, and $X = \begin{bmatrix} X_1 & \cdots & X_n \end{bmatrix}$, for $x = \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix}' \in R^n$, $\omega = \begin{bmatrix} \omega_1 & \cdots & \omega_n \end{bmatrix}' \in R^n$, Ω is a $n \times n$ positive definite matrix.

A.6 Relative standard deviation

Let y_i denote the natural logarithm of Y_i, so $y_i = \ln Y_i$. It is assumed that all selling prices $Y_i > 0$. The model is given by

$$y = X\beta + \varepsilon, \tag{A.1}$$

and $\varepsilon \sim N(0, \sigma^2 I_n)$.

Because of the logarithmic specification of the dependent variable, the standard deviation can be interpreted as a relative standard deviation. Let e denote the vector of residuals, so $e = y - X\widehat{\beta}$, with $\widehat{\beta}$ the Ordinary Least Squares (OLS) estimator of β. Let M_i denote the model value, so $M_i = \exp(X_i\widehat{\beta})$, then

$$y_i - X_i\widehat{\beta} = \ln Y_i - \ln M_i = \ln(1 + \frac{Y_i - M_i}{M_i}) \approx \frac{Y_i - M_i}{M_i},$$

due to the fact that $\ln(1 + \varepsilon) \simeq \varepsilon$, for small ε. If the residuals are not too big, they can be interpreted as relative errors, so the standard deviation from the residuals can be interpreted as a relative standard deviation.

In the standard linear model the residual sum of squares, and hence the standard deviation, is minimized. This means that in the logarithmic specification of the dependent variable the relative errors $(Y - M)/M$ are approximately minimized. If the dependent variable is the selling price, the absolute errors $(Y - M)$ are minimized. In the first case an error of €10.000 on a selling price of €100.000 has a greater impact on the standard deviation than an error of €10.000 on a selling price of €1.000.000. In the last case both errors have the same impact on the standard deviation of the residuals.

In (A.1), $\mathrm{E}\left[y_i - X_i\widehat{\beta}|y, \sigma^2\right] = 0$. The variance is given by $\mathrm{Var}(y_i - X_i\widehat{\beta}|y, \sigma^2) = \tau_i$, with $\tau_i = X_i \mathrm{Var}(\widehat{\beta})X_i' + \sigma^2$, and $\mathrm{Var}(\widehat{\beta}) = \sigma^2(X'X)^{-1}$. The exponent of the model residuals are of interest. The expectation, and variance are given by[1]

$$\begin{aligned}
\mathrm{E}\left[\exp\left\{y_i - X_i\widehat{\beta}\right\}|y, \sigma^2\right] &= \exp\left(\tau_i/2\right), \\
\mathrm{Var}\left(\exp\left\{y_i - X_i\widehat{\beta}\right\}|y, \sigma^2\right) &= \exp\tau_i\left(\exp\tau_i - 1\right).
\end{aligned}$$

Even in the case that $\mathrm{Var}(\widehat{\beta}) = 0$, the expectation is greater than 1, because in general σ^2 is not zero. For example, if $\sigma = 0.15$, $\sigma^2 = 0.0225$, and $\exp\{\sigma^2/2\} \simeq 1.01$. So, a standard deviation of 0.15 leads to over valuation of about 1 percent. So, in order to obtain an expected value of 1, for the ratio between actual and model value all model values can be corrected by a factor $\exp(-\tau_i/2)$.

[1] In general, if $y \sim N(\mu, \sigma^2)$ and $Y = \exp(y)$, then $\mathrm{E}[Y] = \exp\left(\mu + \sigma^2/2\right)$, and $\mathrm{Var}(Y) = \exp(2\mu + \sigma^2)\left(\exp\sigma^2 - 1\right)$. So, in y the expectation and mode coincide, in Y the mode is smaller than the expectation.

A.7 Estimation of multiplicative/additive model

The model (6.2) cannot be estimated by OLS, because it is nonlinear in β. It is quite easy to linearize (6.2) by using the approximation $\ln(1+\varepsilon) \simeq \varepsilon$, for small ε. Define $x(j) = \sum_{i=1}^{k} x_{ij}\beta_i$, and $x^*(j) = \sum_{i=1}^{k} x_{ij}\beta_i^*$ for some β^*, with $\beta_1^* = 1$. So the index j denotes observation j. We can write $\alpha \ln x(j)$ as

$$\begin{aligned}
\alpha \ln x(j) &= \alpha \left[\ln x^*(j) + \ln \left(1 + \sum_{i=2}^{k} \frac{\beta_i - \beta_i^*}{x^*(j)} x_{ij} \right) \right] \\
&\simeq \alpha \left[\ln x^*(j) + \sum_{i=2}^{k} \frac{\beta_i - \beta_i^*}{x^*(j)} x_{ij} \right] \\
&= \alpha \left(\ln x^*(j) - \frac{x^*(j) - x_{1j}}{x^*(j)} \right) + \sum_{i=2}^{k} \alpha\beta_i \frac{x_{ij}}{x^*(j)}.
\end{aligned}$$

So, model (6.2) can be approximated by

$$y_j = \alpha \left(\ln x^*(j) - \frac{x^*(j) - x_1}{x^*(j)} \right) + \sum_{i=2}^{k} \theta_i \frac{x_{ij}}{x^*(j)} + (Z\delta)_j + \varepsilon_j, \tag{A.2}$$

with $\theta_i = \alpha\beta_i$, for $i = 2, \ldots, k$. This model can be estimated as follows:

1. Choose some β^* such that $|\beta_i - \beta_i^*|$ is small,

2. Calculate x^*,

3. Estimate (A.2) by OLS, this provides estimates $\widehat{\alpha}$, and $\widehat{\theta}_i$, so $\widehat{\beta}_i = \widehat{\theta}_i/\widehat{\alpha}$.

4. Substitute β^* with $\widehat{\beta}$, and repeat $1-3$, until $|\beta_i - \beta_i^*| \simeq 0$.

In general this process will converge quickly. A more general approach is provided by Gauss-Newton regression, see for example Davidson and MacKinnon (1993). Consider the model $y = x(\beta) + \varepsilon$, with $x(\beta)$ some nonlinear function in β. Let $\dot{x}(\beta) = \partial x(\beta)/\partial\beta$. The first order Taylor expansion of this model around β^* is provided by

$$\begin{aligned}
y &\simeq x(\beta^*) + \dot{x}(\beta^*)(\beta - \beta^*) + \varepsilon, \\
y - x(\beta^*) &= \dot{x}(\beta^*)b + \varepsilon.
\end{aligned} \tag{A.3}$$

For observation j in (6.2) without $Z\delta$, $\dot{x}_j(\beta)$ is provided by

$$\dot{x}_j(\beta)' = \begin{pmatrix} \ln(X\beta) \\ \alpha\beta_2(\sum_i x_{ij}\beta_i)^{-1}x_{2j} \\ \vdots \\ \alpha\beta_k(\sum_i x_{ij}\beta_i)^{-1}x_{kj} \end{pmatrix}.$$

(A.3) can be estimated by OLS. The estimate of b must equal 0. The OLS must be done recursively, and using gradients and Hessians can speed up convergence, see Davidson and MacKinnon (1993, ch. 6.8). Stop criteria are based on t-statistics.

Bibliography

Andrews, D. W. K. 1994. "The Large Sample Correspondence Between Classical Hypothesis Tests and Bayesian Posterior Odds Tests." *Econometrica* 62:1207–1232.

Anselin, L. 1988. *Spatial Econometrics*. Dordrecht: Kluwer Academic Publishers.

Barndorff-Nielsen, O. 1983. "On a Formula of the Maximum Likelihood Estimator." *Biometrika* 70:343–365.

Bauwens, L., M. Lubrano, and J.-F. Richard. 1999. *Bayesian Inference in Dynamic Econometric Models*. Oxford University Press, Oxford.

Berger, J. O. 2003. "Could Fisher, Jeffreys and Neyman Have Agreed on Testing?" *Statistical Science* 18:1–32.

Berger, J. O., and J. M. Bernardo. 1992. "On the Development of Reference Priors (with Discussion)." In *Bayesian Statistics 4*, edited by J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, 25–60. Oxford University Press, Oxford.

Bernardo, J. M., and A. F. M. Smith. 1994. *Bayesian Theory*. John Wiley, New York.

Bhargava, A. 1986. "On the Theory of Testing for Unit Roots in Observed Time Series." *Review of Economic Studies* 53:137–160.

Box, G. E. P., and G. C. Taio. 1973. *Bayesian Inferenec in Statistical Analysis*. Reading, MA: Addison-Wesley.

Bryk, A. S., and S. W. Raudenbush. 1992. *Hierarchical Linear Models. Applications and Data Analysis Methods*. Newbury Park, CA: Sage Publications.

Can, A. 1992. "Specification and Estimation of Hedonic Price Models." *Regional Science and Urban Economics* 22:453–474.

Carter, C. K., and R. Kohn. 1994. "On Gibbs Sampling for State Space Models." *Biometrika* 81:541–553.

Case, B., and J. M. Quigley. 1991. "The Dynamics of Real Estate Prices." *Review of Economics and Statistics* 73:50–58.

Cooper, D. M., and R. Thompson. 1977. "A Note on the Estimation of the Parameters of the Autoregressive-Moving Average Process." *Biometrika* 64:625–628.

Cox, D. R., and N. Reid. 1987. "Parameter Orthogonality and Approximate Conditional Inference." *Journal of the Royal Statistical Society B* 55:1–39.

———. 1993. "A Note on the Calculation of Adjusted Profile Likelihood." *Journal of the Royal Statistical Society B* 55:467–471.

Davidson, R., and J. G. MacKinnon. 1993. *Estimation and Inference in Econometrics*. Oxford University Press, Oxford.

De Jong, P. 1988. "The Likelihood for a State-Space Model." *Biometrika* 75:165–169.

———. 1991a. "The Diffuse Kalman Filter." *The Annals of Statistics* 2:1073–1083.

———. 1991b. "Stable Algorithms for the State Space Model." *Journal of Time Series Analysis* 12:143–157.

De Jong, P., and S. Chu-Chun Lin. 1994. "Stationary and Non-Stationary State Space Models." *Journal of Time Series Analysis* 15:151–166.

De Vos, A. F. 1998. "Fair and Predictive Bayes Factors for Comparison of Regression Models." Technical Report, Vrije Universiteit Amsterdam, Department of Economics and Econometrics.

De Vos, A. F., and H. R. Merkus. 1992. "Forecasting, Backcasting and Smoothing Algorithms." Technical Report, Vrije Universiteit Amsterdam, Department of Economics and Econometrics.

Dickey, D. A., and W. A. Fuller. 1981. "Likelihood Ratio Tests for Autoregressive Time Series with a Unit Root." *Econometrica* 49:1057–1072.

Dickey, D. A., and W. A Fuller. 1979. "Unit Roots in Time Series Models: Tests and Implications." *Journal of the American Statistical Association* 74:427–431.

Dubin, R. A. 1992. "Spatial Autocorrelation and Neighborhood Quality." *Regional Science and Urban Economics* 22:433–452.

———. 1998. "Predicting House Prices Using Multiple Listings Data." *Journal of Real Estate and Economics* 17:35–59.

Dufour, J.-M., and M. King. 1991. "Optimal Invariant Tests for the Autocorrelation Coefficient in Linear Regression with Stationary or Nonsstationary AR(1) Errors." *Journal of Econometrics* 47:115–143.

Durbin, J., and S. J. Koopman. 2001. *Times Series Analysis by State Space Methods*. Oxford University Press, Oxford.

Elliott, G. E. 1999. "Efficient Tests for a Unit Root When the Initial Observation is Drawn from its Unconditional Distribution." *International Economic Review* 40:767–783.

Elliott, G. E., T. J. Rothenberg, and J. H. Stock. 1996. "Efficient Tests for an Autoregressive Unit Root." *Econometrica* 64:813–836.

Fernández, C., and M. F. J. Steel. 1999. "Reference Priors for the General Location-Scale Model." *Statististics and Probability Letters* 43:377–384.

Fleming, M. C., and J. G. Nellis. 1992. "Development of Standardized Indices for Measuring House Price Inflation Incorporating Physical and Locational Characteristics." *Applied Economics* 24:1067–1085.

Francke, M. K. 1996. "De Waarde van omgevingskenmerken." In *Waardebepaling Vastgoed. Enkele Actuele Ontwikkelingen*, edited by L.B. Uittenbogaard and G. A. Vos, 74–88. Stichting voor Beleggings- en Vastgoedkunde.

Francke, M. K., and A. F. de Vos. 2000. "Efficient Computation of Hierarchical Trends." *Journal of Business and Economic Statistics* 18:51–57.

———. 2006. "Marginal Likelihood and Unit Roots." *Journal of Econometrics*. to appear.

Francke, M. K., and G. A. Vos. 2004. "The Hierarchical Trend Model for Property Valuation and Local Price Indices." *Journal of Real Estate Finance and Economics* 28:179–208.

Frühwirth-Schnatter, S. 1994. "Data Augmentation and Dynamic Linear Models." *Journal of Time Series Analysis* 15:183–202.

Fuller, W. A. 1976. *Introduction to Statistical Time Series*. John Wiley, New York.

Gelfland, A. E., S. K. Ghosh, J. R. Knight, and C. F. Sirmans. 1998. "Spatio-Temporal Modeling of Residential Sales Data." *Journal of Business and Economic Statistics* 16:312–321.

Ghosh, M., and J. Heo. 2003. "Default Bayesian Priors for Regression Models with First-Order Autoregressive Residuals." *Journal of Time Series Analysis* 24:269–282.

Goldstein, H. 1995. *Kendall's Advanced Theory of Statistics, Vol. 3, Multilevel Statistical Models*. London: Arnold.

Halvorsen, R., and H. Pollakowski. 1981. "Choice of the Functional Form for Hedonic Price Equations." *Journal of Urban Economics* 10:37–49.

Harvey, A. 1989. *Forecasting Structural Time Series Models and the Kalman Filter*. Cambridge University Press, Cambridge.

———. 2005. "A Unified Approach to Testing for Stationarity and Unit Roots." In *Identification and Inference for Econometric Models. A Festschrift for Tom Rothenberg*, edited by D. Andrews, J. Powell, P. Ruud, and J. Stock, 403–425. Cambridge University Press, Cambridge.

Harville, D. A. 1974. "Bayesian Inference for Variance Components Using Only Error Contrast." *Biometrika* 61:383–385.

Kalbfleisch, J. D., and D. A. Sprott. 1970. "Application of Likelihood Methods to Models Involving Large Numbers of Parameters." *Journal of the Royal Statistical Society B* 32:175–208.

King, M. L. 1980. "Robust Tests for Spherical Symmetry and their Application to Least Squares Regression." *The Annals of Statistics* 8:1630–1638.

Kitagawa, G. 1994. "The Two-Filter Formula for Smoothing and Implementation of the Gaussian-Sum Smoother." *Annals of the Institute of Statistical Mathematics* 46:605–623.

Knight, J. R., J. Dombrow, and C. F. Sirmans. 1995. "A Varying Parameters Approach to Constructing House Price Indexes." *Real Estate Economics* 23:187–205.

Koopman, S. J. 1992. *Diagnostic Checking and Intra-Daily Effects in Time Series Models*. Amsterdam: Thesis Publishers.

———. 1997. "Exact Initial Kalman Filtering and Smoothing for Nonstationary Time Series Models." *Journal of the American Statistical Assocation* 92:1630–1638.

Kuo, B. S. 1999. "Asymptotics of ML Estimator for Regression Models with a Stochastic Trend Component." *Econometric Theory* 15:24–29.

Lee, T., and D. A. Dickey. 2004. "Limiting Distributions of Unconditional Maximum Likelihood Unit Root Test Statistics in Seasonal Time-Series Models." *Journal of Time Series Analysis*, pp. 551–561.

Lehmann, E. L. 1986. *Testing Statistical Hypotheses*. 2. John Wiley, New York.

Levenbach, H. 1972. "Estimation of Autoregressive Parameters from a Marginal Likelihood Function." *Biometrika* 59:61–71.

Longford, N. T. 1993. *Random Coefficients Models*. Oxford: Clarendon Press.

Lubrano, M. 1995. "Testing for Unit Roots in a Bayesian Framework." *Journal of Econometrics* 69:81–109.

McCullagh, P., and J. A. Nelder. 1989. *Generalized Linear Models*. 2. London: Chapman & Hall.

Merkus, H. R., D. S. G. Pollock, and A. F. de Vos. 1993. "A Synopsis of the Smoothing Formulae Associated with the Kalman Filter." *Computational Economics* 6:177–200.

Müller, U.K., and G. Elliott. 2003. "Tests for Unit Roots and the Initial Condition." *Econometrica* 71:1269–1286.

Needham, B., M. K. Francke, and P. Bosma. 1997/1998. "How the City of Amsterdam is Using Econometric Modelling to Value Real Estate." *Journal of Property Tax Assessment & Administration* 3:25–46.

Ng, S., and P. Perron. 2001. "Lag Length Selection and the Construction of Unit Root Tests with Good Size and Power." *Econometrica* 71:1519–1554.

O'Hagan, A. 1994. *Kendall's Advanced Theory of Statistics, Vol. 2B, Bayesian Inference.* London: Arnold.

Orford, S. 1999. *Valuing the Built Environment: GIS and House Price Analysis.* Brookfield, VT, U.S.A.: Ashage Publishing Company.

Pace, R. K., Barry R., J. M. Clapp, and M. Rodriquez. 1998. "Spatiotemporal Autoregressive Models of Neighborhood Effects." *Journal of Real Estate Finance and Economics* 17:15–33.

Palmquist, R. B. 1984. "Estimating the Demand for the Characteristics of Housing." *Review of Economics and Statistics* 66:394–404.

Pantula, S., G. Gonzalez-Farias, and W. Fuller. 1994. "A Comparison of Unit-Root Test Criteria." *Journal of Business and Economic Statistics* 12:449–459.

Patterson, H. D., and R. Thompson. 1971. "Recovery of Inter-Block Information When Block Sizes are Unequal." *Biometrika* 58:545–554.

Pere, P. 2003. "AR(1) Models, Unit Roots, and Adjusted Profile Likelihood." *Econometric Theory* 19:885–922.

Phillips, P. C. B. 1991. "To Criticize the Critics: An Objective Bayesian Analysis of Stochastic Trends." *Journal of Applied Econometrics* 6:333–364.

Poirier, D. J. 1995. *Intermediate Statistics and Econometrics.* Cambridge: MIT.

Rahman, S., and M. L. King. 1997. "Marginal-Likelihood Score-Based Tests of Regression Disturbances in the Presence of Nuisance Parameters." *Journal of Econometrics* 82:81–106.

Schmidt, P., and P. C. B. Phillips. 1992. "LM Tests for a Unit Root in the Presence of Deterministic Trends." *Oxford Bulletin of Economics and Statistics* 54:257–287.

Schwann, G. M. 1998. "A Real Estate Price Index for Thin Markets." *Journal of Real Estate Finance and Economics* 16:269–287.

Shephard, N. 1993. "Maximum Likelihood Estimation of Regression Models with Stochastic Trend Components." *Journal of the American Statistical Association* 88:590–595.

———. 1994. "Partial Non-Gaussian State Space." *Biometrika* 81:115–131.

Shephard, N., and A.C. Harvey. 1990. "On the Probability of Estimating a Deterministic Component in the Local Level Model." *Journal of Time Series Analysis* 11:339–347.

Shin, D. W., and W. A. Fuller. 1998. "Unit Root Tests Based on Unconditional Maximum Likelihood Estimation for the Autoregressive Moving Average." *Journal of Time Series Analysis* 19:591–599.

Sims, C. A. 1988. "Bayesian Skepticism on Unit Root Econometrics." *Journal of Economic Dynamics and Control* 12:463–474.

Sims, C. A., and H. Uhlig. 1991. "Understanding Unit Rooters: A Helicopter Tour." *Econometrica* 59:1591–1599.

Smyth, G. K., and A. P. Verbyla. 1996. "A Conditional Likelihood Approach to REML in Generalized Linear Models." *Journal of the Royal Statistical Society B* 58:565–572.

Stone, M., and A. P. Dawid. 1972. "Un-Bayesian Implications of Improper Bayes Inference in Routine Statistical Problems." *Biometrika* 59:369–375.

Tunnicliffe Wilson, G. 1989. "On the Use of Marginal Likelihood in Time Series Model Estimation." *Journal of the Royal Statistical Society B* 51:15–27.

West, M., and J. Harrisson. 1997. *Bayesian Forecasting and Dynamic Models*. 2. New York: Springer-Verlag.

Wolverton, M. L., and J. Senteza. 2000. "Hedonic Estimates of Regional Constant Quality House Prices." *Journal of Propertyy Research* 17:93–108.

Zellner, A. 1971. *An Introduction to Bayesian Inference in Econometrics*. John Wiley, New York.

Samenvatting (Summary)

Marginale aannemelijkheid in toestandsruimte modellen
Theorie en toepassingen

Dit proefschrift gaat over de marginale aannemelijkheid in lineaire toestandsruimte modellen met normaal verdeelde storingen. Veel in de econometrie gebruikte tijdreeksmodellen kunnen worden geschreven als een toestandsruimte model. Een voorbeeld van zo'n model is het in hoofdstuk 5 en 6 beschreven hiërarchische trend model, waarin verkoopprijzen van woningen worden verklaard aan de hand van objectkenmerken, de locatie en het tijdstip van verkoop.

Een toestandsruimte model beschrijft de relatie tussen niet waarneembare toestanden en waarnemingen en bestaat uit een meet- en een toestandsvergelijking. In de toestandsvergelijking wordt beschreven hoe de toestand in een volgende periode afhangt van de toestand in de huidige periode. De meetvergelijking beschrijft de relatie tussen de toestand en de overige variabelen in het model. In beide vergelijkingen wordt verondersteld dat de storingen normaal verdeeld zijn. De meet- en toestandsvergelijking hangen af van onbekende systeemparameters. Als de initiële toestand en de systeemparameters bekend zijn, kan met behulp van het Kalman filter recursief de toestand op tijdstip t worden geschat op basis van de waarnemingen tot en met tijdstip t. Het Kalman filter geeft eveneens een aannemelijkheidfunctie waarmee de meest aannemelijke schatter van de onbekende systeemparameters kan worden bepaald.

In veel toestandsruimte modellen is de initiële toestand, de toestand op tijdstip $t = 1$, onbekend. Het diffuse Kalman filter houdt hiermee rekening en geeft conditioneel op de systeemparameters voor ieder tijdstip een schatting van de toestand, inclusief de initiële toestand. Echter, de aannemelijkheidfunctie is afhankelijk van hoe de initiële toestand wordt behandeld. De initiële toestand kan worden beschouwd als een vaste onbekende waarde of als een random variabele met een diffuse prior (voorverdeling), leidend tot respectievelijk de profile en diffuse aannemelijkheidfunctie, zie De Jong (1991a). De schattingsresultaten van de systeemparameters zijn afhankelijk van de aannemelijkheidfunctie die wordt gebruikt. Dit proefschrift gaat

nader in op het aannemelijkheidbegrip in toestandsruimte modellen en legt een verbinding tussen klassieke en Bayesiaanse statistische literatuur.

In de Kalman filter literatuur is nauwelijks enige motivatie te vinden welke aannemelijkheidfunctie gebruikt zou moeten worden. Uitzonderingen zijn studies van Shephard and Harvey (1990) en Shephard (1993), waarin voor specifieke modellen wordt aangetoond dat de diffuse aannemelijkheidfunctie tot betere schattingsresultaten leidt.

Een rechtvaardiging voor het gebruik van de diffuse aannemelijkheidfunctie kan worden gevonden in de literatuur uit de jaren zeventig over klassieke marginale aannemelijkheid in het regressiemodel, waarin de variantie covariantie matrix afhangt van onbekende parameters. De regressie- en schaalparameters zijn nuisance (hinderlijke, overtollige) parameters. Een toestandsruimte model kan gezien worden als een bijzondere vorm van een regressiemodel met een variantie covariantie matrix die afhankelijk is van de specificatie van de toestandsvergelijking. De regressieparameters corresponderen met de onbekende initiële toestand en de variantie covariantie matrix is afhankelijk van de systeemparameters.

De diffuse aannemelijkheidfunctie is proportioneel met de klassieke marginale aannemelijkheidfunctie. Het verschil is een term die in de meeste gevallen niet afhangt van de onbekende parameters en dan ook irrelevant is.

In hoofdstuk 2 wordt allereerst het begrip klassieke marginale aannemelijkheid nader omschreven. De klassieke marginale aannemelijkheid komt overeen met de verdeling van een maximaal invariante transformatie van de data, zodanig dat de getransformeerde data onafhankelijk zijn van de regressie- en schaalparameters. De claim is dat het complement van de marginale aannemelijkheid geen informatie bevat over de parameters uit de variantie covariantie matrix en dus buiten beschouwing moet worden gelaten voor het trekken van conclusies ten aanzien van de parameters uit de variantie covariantie matrix. Het verschil met de profile aannemelijkheid, dat is de volle aannemelijkheid waarin de meest aannemelijke schatters voor de regressie- en schaalparameters zijn ingevuld, is een functie die afhangt van de onbekende parameters uit de variantie covariantie matrix.

In het vervolg van dit hoofdstuk wordt uitgebreid ingegaan op een voorbeeld waarin het verschil tussen de profile en marginale aannemelijkheid (en dus de diffuse aannemelijkheid) extreem belangrijk is, namelijk het eenheidswortel probleem in het regressiemodel met autoregressieve storingen. Zowel de exacte specificatie van dit model als de aanname ten aanzien van de initiële waarneming (of toestand in de toestandsruimte formulering) is van essentieel belang voor het trekken van conclusies over de autoregressieve parameter ρ. Het in hoofdstuk 2 gepropageerde model heeft een coherente betekenis voor het verwachte niveau van het proces voor $|\rho| < 1$. De veronderstelde initiële conditie is coherent voor $\rho \uparrow 1$. Voor dit model wordt zowel de profile als de marginale aannemelijkheidfunctie afgeleid. De profile aannemelijkheidfunctie is in dit model gelijk aan 0 voor $\rho = 1$ en kan dus niet worden gebruikt voor aannemelijkheidverhouding toetsen. De marginale aannemelijkheidfunctie echter is eindig en ongelijk aan 0 voor $\rho = 1$ en continu voor $\rho \uparrow 1$.

De marginale aannemelijkheid hangt slechts af van één parameter, namelijk ρ, zodat het Neyman-Pearson lemma de optimale kritieke zone definieert voor iedere vaste alternatieve waarde van ρ ten opzichte van de nulhypothese $\rho = 1$. Tevens kan hiermee de power envelope (omhullend onderscheidend vermogen) worden bepaald. De asymptotische verdeling van de marginale aannemelijkheidverhouding onder de nulhypothese is geëvalueerd in het "local-to-unity" formaat $\gamma = n(1 - \rho)$, waarbij γ een vaste constante is als het aantal waarnemingen n naar oneindig gaat. Het is een functie van meerdere statistische grootheden met gewichten die afhangen van γ, zodat zelfs asymptotisch geen meest uniform meest onderscheidende toets (UMP) bestaat. In het geval van een onbekende γ wordt als toetsingsgrootheid de marginale aannemelijkheidverhouding geëvalueerd in de marginaal meest aannemelijke schatter van γ. De resulterende toetsen zijn krachtiger dan andere uit de literatuur bekende toetsen. De power functie (onderscheidend vermogen als functie van de alternatieve hypothese) van de marginale aannemelijkheidverhouding toetsen valt ook voor kleine steekproeven vrijwel samen met de power envelope, ondanks het feit dat er geen UMP toets bestaat. Een verklaring hiervoor is dat de aannemelijkheidverhouding functie vrijwel monotoon is in een statistische grootheid.

De toets kan worden aangepast om rekening te houden met complexere covariantie structuren, namelijk $u_{t+1} = \rho u_t + v_t$, waarbij v_t in plaats van ongecorreleerd, serieel gecorreleerd is. Eén van de mogelijkheden is het gebruiken van standaard aannemelijkheidmethoden voor het schatten en selecteren van modellen, gebaseerd op de marginale aannemelijkheid. Een andere mogelijkheid is een aangepaste marginale aannemelijkheidverhouding toets, waarbij de aanpassing gebaseerd is op de asymptotische verdeling van de marginale aannemelijkheidverhouding onder seriële correlatie. Het verschil met de asymptotische verdeling zonder seriële correlatie hangt af van de verhouding tussen de lange termijn en de onconditionele variantie van v_t. Deze verhouding kan bijvoorbeeld consistent geschat worden met behulp van de kleinste kwadraten methode. Een simulatiestudie waarbij v_t een eerste orde voortschrijdend gemiddelde proces volgt, laat zien dat de power functies van de aangepaste toetsen andere toetsen overtreffen.

Ook in de Bayesiaanse statistiek wordt de term marginale aannemelijkheid gebruikt. Hoofdstuk 3 gaat nader in op het Bayesiaanse marginale aannemelijkheidbegrip en de relatie met de klassieke marginale aannemelijkheid binnen het lineaire model waarin de regressie- en plaatsparameters als nuisance worden beschouwd. De Bayesiaanse marginale aannemelijkheid is de aannemelijkheid die wordt verkregen door het "uitintegreren" van de nuisance parameters. De aldus verkregen marginale aannemelijkheid is afhankelijk van de priorverdeling van de nuisance parameters conditioneel op de overige parameters. In hoofdstuk 3 wordt afgeleid dat de Bayesiaanse marginale aannemelijkheid proportioneel is met de klassieke marginale aannemelijkheid bij het gebruik van de onafhankelijke Jeffreys' prior. Deze prior wijkt af van de prior die volgt uit Jeffreys' regel. Beide niet informatieve voorverdelingen zijn "improper", zodat de Bayesiaanse marginale aannemelijkheid gedefinieerd is op een constante na. Het gebruik van de klassieke marginale aannemelijkheid voorkomt het probleem van niet goed gedefinieerde voorverdelingen en marginale aannemelijkheidfuncties.

Op het gebied van toetsen bestaan er grote verschillen tussen de Bayesiaanse en klassieke aanpak. Echter, in het geval van een monotone marginale aannemelijkheidverhouding en een marginale aannemelijkheidfunctie die slechts afhangt van één parameter, zijn er grote overeenkomsten bij eenzijdige toetsen. Bij het gebruik van goed gedefinieerde priors, d.w.z. priors die een kansmassa van 1 hebben, is het enige verschil tussen de a posteriori kansverhouding toets (Bayes factor) en de klassieke marginale aannemelijkheidverhouding toets de omvang (α) van de toets. Hoewel het niet zo gebruikelijk is binnen het Bayesiaanse paradigma, kan ook voor de Bayes factor een p-waarde worden bepaald. Als gevolg van de monotoniciteit van de marginale aannemelijkheidverhouding is deze p-waarde onafhankelijk van de prior. Omdat de klassieke marginale aannemelijkheidverhouding toets een uniform meest onderscheidende invariante toets is, geldt dit ook voor de toets op basis van de Bayes factor.

Beide resultaten, de proportionaliteit van de klassieke en Bayesiaanse marginale aannemelijkheid en hetzelfde gebruik van de data in het toetsen, worden toegepast op het eenheidswortel voorbeeld, waarin de marginale aannemelijkheidverhouding bij benadering monotoon is. In de Bayesiaanse literatuur is voor dit model een aantal niet-informatieve priors voor de nuisance parameters voorgesteld die in combinatie met een goed gedefinieerde prior voor ρ leiden tot een naverdeling die gelijk is aan 0 voor $\rho = 1$. Het gebruik van de onafhankelijke Jeffreys' prior, of het direct gebruiken van de klassiek marginale aannemelijkheid, voorkomt dit probleem. Voor twee verschillende priors voor ρ wordt aangetoond dat de power functie van de bijbehorende a posteriori kansverhouding toets vrijwel samenvalt met de power envelope. Het enige verschil tussen de klassieke en Bayesiaanse toetsen is de grootte van de toets. Alleen voor priors die worden geformuleerd in termen van $\gamma = n(1 - \rho)$ in plaats van ρ, is overeenkomst mogelijk in de grootte van de toets voor alle n.

In hoofdstuk 4 komt het schatten en toetsen van systeemparameters in een toestandsruimte model met diffuse intiële condities aan de orde. De verschillende aannemelijkheidconcepten, de profile, diffuse en marginale aannemelijkheid, worden met elkaar vergeleken. Het gebruik van de marginale aannemelijkheid en met enige voorzichtigheid de diffuse, is te prefereren boven het gebruik van de profile aannemelijkheid. Een overtuigend voorbeeld hiervan is het eenheidswortel voorbeeld. Een efficiënte manier om de marginale aannemelijkheid te berekenen, is een aangepaste versie van het diffuse Kalman filter.

In de meeste gevallen is de marginale aannemelijkheid proportioneel met de diffuse aannemelijkheid. Echter, in tegenstelling tot de marginale aannemelijkheid is de diffuse aannemelijkheid afhankelijk van de specifieke formulering van het toestandsruimte model. Voor sommige modellen is het verschil tussen de diffuse aannemelijkheidfuncties in verschillende specificaties van hetzelfde model afhankelijk van de systeemparameters. De marginale aannemelijkheid is daarom te verkiezen boven de diffuse aannemelijkheid.

Formeel gezien mogen de marginale en diffuse aannemelijkheid niet gebruikt worden voor het schatten van systeemparameters in niet-lineaire modellen, omdat deze gebaseerd zijn op parameter afhankelijke transformaties van de data. Er is daarom een alternatieve schattings-

methode gegeven, die is gebaseerd op een eerste orde benadering van het niet-lineaire model.

De marginale aannemelijkheid en in het bijzonder de diffuse aannemelijkheid kan niet gebruikt worden voor "goodness of fit" toetsen, zoals het Akaike en het Bayesiaanse Informatie Criterium, omdat voor verschillende modellen verschillende transformaties van de data worden gebruikt. Voor geneste modellen wordt een alternatieve procedure gegeven die is gebaseerd op de marginale aannemelijkheid. In een simulatievoorbeeld wordt aangetoond dat deze procedure tot betere resultaten leidt dan een procedure die is gebaseerd op de profile aannemelijkheid.

De hoofdstukken 5 en 6 bevatten een voorbeeld van een toestandsruimte model dat wordt gebruikt voor het waarderen van woningen. In dit hiërarchische trend model worden verkoopprijzen van woningen verklaard aan de hand van kenmerken en spelen een algemene trend en cluster trends een rol. De algemene en de cluster trends als afwijkingen van de algemene trend, zijn gemodelleerd als stochastische trends, bijvoorbeeld een random walk. Het model is geschreven in het formaat van een toestandsruimte model, waarin de initiële toestand onbekend is.

Hoofdstuk 5 beschrijft voor dit model een efficiënte schattingsprocedure. De structuur van herhaalde metingen maakt het mogelijk om het model te ontbinden in twee delen, namelijk een model met gemiddelden per cluster per tijdstip en een model met afwijkingen van de gemiddelden per cluster per tijdstip. Het laatste model kan simpel geschat worden met behulp van de kleinste kwadraten methode. Dit levert een prior op die gebruikt wordt in het model met de gemiddelden. Deze aanpak maakt gebruik van de Bayesiaanse interpretatie van het Kalman filter en vormt voor dit model een alternatief voor het diffuse Kalman filter. De totale aannemelijkheid, die gebruikt kan worden voor het maximaliseren naar de systeemparameters, is eenvoudig het product van de aannemelijkheidfuncties van de twee afzonderlijke modellen. Een volledig Bayesiaans alternatief voor het schatten van de onbekende parameters en de toestandsvectoren, is de "Gibbs sampler", waarin op een efficiënte wijze gebruik gemaakt kan worden van de output van het Kalman filter, zie bijvoorbeeld Frühwirth-Schnatter (1994).

Het hiërarchische trend model wordt toegepast op een dataverzameling van Dienst Belastingen Gemeente Amsterdam met verkoopprijzen en karakteristieken van ruim 12.000 stapelwoningen in Amsterdam over een periode van ruim tien jaar. De prijsontwikkeling volgt een algemeen patroon, maar kan variëren per gebied. De clusters worden daarom gevormd door gebieden in Amsterdam. In dit voorbeeld wijken de schattingen van de coëfficiënten uit de regressie op de afwijkingen van de gemiddelden per tijdstip per cluster nauwelijks af van de uiteindelijke schattingen van het Kalman filter. De functionele vorm voor de verklarende variabelen kan dus eenvoudig gebaseerd worden op de waarnemingen in afwijking van de gemiddelden.

Hoofdstuk 6 gaat verder in op toepassingen van het hiërarchische trend model. Dit model kan gezien worden als een hedonisch prijsmodel waarin verkoopprijzen worden verklaard aan de hand van de kenmerken van de woningen. Daarnaast wordt in het bijzonder rekening gehouden met de ruimtelijke en temporele aspecten van de verkoopprijzen. Er worden twee verschillende soorten clusters onderscheiden, namelijk clusters gebaseerd op gebieden en woningtyperingen.

Het model wordt in de praktijk gebruikt voor de waardebepaling van woningen en het vaststellen van prijsontwikkelingen op lokaal niveau in het kader van de Wet Waardering Onroerende Zaken. In dit hoofdstuk wordt het model toegepast op twee dataverzamelingen van de Nederlandse Vereniging van Makelaren (NVM), namelijk die voor de woningmarkt in Amsterdam en Breda. Voor beide gebieden worden de schattingsresultaten getoond. De uit het model verkregen prijsontwikkelingen worden voorts vergeleken met twee andere methoden. De eerste methode, die wordt gebruikt door de NVM, is op basis van gewogen mediane verkoopprijzen. De tweede methode maakt gebruik van een standaard hedonisch regressiemodel met dummy's per periode. Uit de vergelijking van de resultaten blijkt dat de prijsontwikkelingen op basis van het hiërarchische trend model betrouwbaarder, gedetailleerder en meer up-to-date zijn dan die op basis van de andere twee methoden. Dit geldt in het bijzonder voor clusters met relatief weinig waarnemingen.

Dankwoord (Acknowledgements)

Twaalf jaar geleden ben ik begonnen met mijn promotieonderzoek. Soms is het beter om niet alles vooraf te weten, want in dat geval was ik vrijwel zeker niet aan dit avontuur begonnen en achteraf ben ik blij dat ik het heb gedaan.

Op 1 maart 1994 liep mijn stage bij Dienst Belastingen Gemeente Amsterdam vrijwel geruisloos over in een baan bij dezelfde dienst met daarbij de mogelijkheid om voor één dag per week promotieonderzoek te doen. Werk en promotieonderzoek lagen aanvankelijk in elkaars verlengde, namelijk het ontwikkelen van waarderingsmodellen voor woningen. Deze modellen zijn inderdaad ontwikkeld en beproefd in de praktijk, maar gaandeweg het traject zijn de paden van onderzoek en werk meer en meer uit elkaar gaan lopen. De aandacht van het promotieonderzoek is geleidelijk aan verschoven van toegepast onderzoek naar theoretische econometrische problemen, die ogenschijnlijk niet zo veel meer van doen hebben met de oorspronkelijke toepassing. Toch heeft de wisselwerking tussen theorie en praktijk mij in de loop der tijd juist gestimuleerd.

Graag wil ik een aantal mensen bedanken die betrokken zijn geweest bij de totstandkoming van dit proefschrift. Allereerst mijn copromotor, dr. Aart de Vos, die mij heeft overgehaald om aan dit onderzoek te beginnen. En in het bijzonder voor de jarenlange strijdbare samenwerking en vriendschap. In eerste instantie zijn we het niet vaak met elkaar eens, maar uiteindelijk komen we er toch altijd uit. Beste Aart, ik heb veel van jouw creatieve ideeën geleerd, die je vol enthousiasme met me deelde. Ook wil ik Annemarie en jou hartelijk bedanken voor het geven van het laatste zetje voor het afronden van mijn proefschrift door het productieve en aangename verblijf in Collabassa.

Mijn promotor prof.dr. Siem Jan Koopman dank ik voor zijn kundige, adequate en prettige begeleiding gedurende de laatste periode van het promotieonderzoek.

Drs. Gerjan Vos ben ik dankbaar voor het mij introduceren in de wetenschappelijke wereld van het vastgoed en voor de aangename samenwerking in het onderzoek naar hedonische prijsindices.

De leden van de leescomissie, prof.dr. Peter Boswijk, dr. Foort Hamelink, prof.dr. Andrew Harvey en prof.dr. Michel Lubrano ben ik erkentelijk voor het zorgvuldig bestuderen van het manuscript en het aandragen van waardevolle suggesties.

Joop Dorrepaal, Willem Bruul en Pieter Bosma wil ik bedanken voor het bieden van de

mogelijkheid om promotieonderzoek uit te voeren bij de Dienst Belastingen Gemeente Amsterdam. Ook wil ik mijn toenmalige collega's van Vastgoed en in het bijzonder van de afdeling Modelontwikkeling en Marktanalyse danken voor de plezierige werksfeer.

Mijn huidige collega's van OrtaX dank ik voor de goede samenwerking en het verder ontwikkelen van de waarderingsmodellen.

Raymond Havekes, mededirecteur en paranimf, dank ik voor de jarenlange intensieve en hechte samenwerking. Na je een paar jaar uit het oog te hebben verloren, kruisten onze wegen elkaar weer bij de oprichting van OrtaX.

Kai Ming Lee ben ik erkentelijk voor het geheel belangeloos helpen bij het oplossen van allerlei LaTex problemen.

Theo Goverts, paranimf, dank ik voor zijn jarenlange vriendschap. Mijn familie en vrienden ben ik dankbaar voor de jarenlange steun, die veel verder gaat en meer omvat dan dit proefschrift. Tenslotte dank ik mijn ouders voor hun voortdurende waardering en ondersteuning.

Amsterdam, januari 2006

Curriculum Vitae

Marc K. Francke was born on March 8, 1970 in Middelburg. He graduated in 1994 in Econometrics at the Vrije Universiteit Amsterdam. During 1994 - 2000 he worked at the Amsterdam Tax Authorities office, where he developed models for mass appraisal of real estate. In 2001 he was co-founder of OrtaX, a company specialized in mass appraisal for local government and housing corporations.